

Google Sitemap Generator Help

Index

Google Sitemap Generator Help	1
Index	2
License	3
1 Sitemap Introduction.....	7
1.1 Common Sitemap Introduction	7
1.2 News Sitemap Introduction (available for News site only).....	7
2 Google Sitemap Generator Introduction	9
2.1 Basic Introduction	9
2.2 Mechanism and Limitation.....	9
3 Install Google Sitemap Generator.....	10
3.1 System Requirement.....	10
3.2 Installation Steps	11
3.3 Uninstallation Steps.....	14
4 Configure Google Sitemap Generator.....	19
4.1 Introduction	19
4.2 Site setting describe.....	26
4.2.1 Global Site Setting:	26
4.2.2 Normal Site Setting:	27
4.2.3 Web Sitemap Setting:	29
4.2.4 News Sitemap Setting:	31
4.2.5 Video Sitemap Setting:.....	32
4.2.6 Mobile Sitemap Setting:.....	33
4.2.7 Code Search Sitemap Setting:	34
4.2.8 Blog Search Ping Setting.....	35
4.2.9 Runtime Info:	36
4.3 FAQ	39
4.3.1 How to set the setting port?.....	39
4.3.2 How to enable the Web and other sitemaps?	39
4.3.3 How to limit the disk&mem space using?.....	40
4.3.4 I don't know what's the valid value to input, what can I do?.....	41
4.3.5 What are the URL pattern language rules?.....	41
5 Trouble Shooting.....	42
6 Contact	42

License

Copyright 2008 Google Inc.

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License.

You may obtain a copy of the License at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.

See the License for the specific language governing permissions and limitations under the License.

Apache License

Version 2.0, January 2004

<http://www.apache.org/licenses/>

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

1. Definitions.

"License" shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

"Licensor" shall mean the copyright owner or entity authorized by the copyright owner that is granting the License.

"Legal Entity" shall mean the union of the acting entity and all other entities that control, are controlled by, or are under common control with that entity. For the purposes of this definition, "control" means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

"You" (or "Your") shall mean an individual or Legal Entity exercising permissions granted by this License.

"Source" form shall mean the preferred form for making modifications, including but not limited to software source code, documentation source, and configuration files.

"Object" form shall mean any form resulting from mechanical transformation or translation of a Source form, including but not limited to compiled object code, generated documentation, and conversions to other media types.

"Work" shall mean the work of authorship, whether in Source or Object form, made available under the License, as indicated by a copyright notice that is included in or attached to the work (an example is provided in the Appendix below).

"Derivative Works" shall mean any work, whether in Source or Object form, that is based on (or derived from) the Work and for which the editorial revisions, annotations, elaborations, or other modifications represent, as a whole, an original work of authorship. For the purposes of this License, Derivative Works shall not include works that remain separable from, or merely link (or bind by name) to the interfaces of, the Work and Derivative Works thereof.

"Contribution" shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright owner or by an individual or Legal Entity authorized to submit on behalf of the copyright owner. For the purposes of this definition, "submitted" means any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems that are managed by, or on behalf of, the Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as "Not a Contribution."

"Contributor" shall mean Licensor and any individual or Legal Entity on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

2. Grant of Copyright License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

3. Grant of Patent License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was submitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

4. Redistribution. You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

- a. You must give any other recipients of the Work or Derivative Works a copy of this License; and
- b. You must cause any modified files to carry prominent notices stating that You changed the files; and
- c. You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
- d. If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.

You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions. Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and customary use in describing the origin of the Work and reproducing the content of the NOTICE file.

7. Disclaimer of Warranty. Unless required by applicable law or agreed to in writing, Licensor provides the Work (and each Contributor provides its Contributions) on an

"AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied, including, without limitation, any warranties or conditions of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A PARTICULAR PURPOSE. You are solely responsible for determining the appropriateness of using or redistributing the Work and assume any risks associated with Your exercise of permissions under this License.

8. Limitation of Liability. In no event and under no legal theory, whether in tort (including negligence), contract, or otherwise, unless required by applicable law (such as deliberate and grossly negligent acts) or agreed to in writing, shall any Contributor be liable to You for damages, including any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or out of the use or inability to use the Work (including but not limited to damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses), even if such Contributor has been advised of the possibility of such damages.

9. Accepting Warranty or Additional Liability. While redistributing the Work or Derivative Works thereof, You may choose to offer, and charge a fee for, acceptance of support, warranty, indemnity, or other liability obligations and/or rights consistent with this License. However, in accepting such obligations, You may act only on Your own behalf and on Your sole responsibility, not on behalf of any other Contributor, and only if You agree to indemnify, defend, and hold each Contributor harmless for any liability incurred by, or claims asserted against, such Contributor by reason of your accepting any such warranty or additional liability.

END OF TERMS AND CONDITIONS

1 Sitemap Introduction

1.1 Web Sitemap Introduction

XML Sitemaps - usually called Sitemaps, with a capital S - are a way for you to give Google information about your site.

In its simplest terms, a Sitemap is a list of the pages on your website. Creating and submitting a Sitemap helps make sure that Google knows about all the pages on your site, including URLs that may not be discoverable by Google's normal crawling process.

Sitemaps are particularly helpful if:

- ❖ Your site has dynamic content.
- ❖ Your site has pages that aren't easily discovered by Googlebot during the crawl process - for example, pages featuring rich AJAX or Flash.
- ❖ Your site is new and has few links to it. (Googlebot crawls the web by following links from one page to another, so if your site isn't well linked, it may be hard for us to discover it.)
- ❖ Your site has a large archive of content pages that are not well linked to each other, or are not linked at all.

The Sitemap Protocol v0.90 is also supported by the search engines of Yahoo! and Microsoft.

For more information about sitemap, you can refer to

<http://www.sitemaps.org/>

<http://www.google.com/support/webmasters/bin/topic.py?topic=8467>.

1.2 News Sitemap Introduction

News sitemap file is an extension of Web sitemap file. It can submit new published news content to Google News Search Engine quickly and automatically.

It is only supported by Google Search Engine up-to-now, and only News web sites that have been authorized by Google can submit News sitemaps. For more information, please see

<http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=42738>.

1.3 Video Sitemap Introduction

Google Video Sitemaps is an extension of the Sitemap protocol that enables you to

publish and syndicate online video content and its relevant metadata to Google in order to make it searchable in the Google Video index. You can use a Video Sitemap to add descriptive information - such as a video's title, description, duration, etc. - that makes it easier for users to find a particular piece of content. When a user finds your video through Google, they will be linked to your hosted environments for the full playback.

For more information, please see

<http://www.google.com/support/webmasters/bin/topic.py?topic=10079>

1.4 Mobile Sitemap Introduction

Google Mobile Sitemaps is an extension of the Sitemap protocol that enables you to submit URLs that serve content for mobile devices into our mobile index. By using Mobile Sitemaps to inform and direct our crawlers, you can expand our coverage to your mobile web and speed up the discovery and addition of pages in your site to our mobile index, which used by Google Mobile Web Search.

For more information, please see

<http://www.google.com/support/webmasters/bin/topic.py?topic=8493>

1.5 Code Search Sitemap Introduction

Google's Code Search helps users find function definitions and sample code by enabling them to search publicly accessible source code hosted on the Internet. You can tell Google about source code on your site by creating and submitting a Code Search Sitemap. A Code Search Sitemap is just like a regular Sitemap, and is submitted in the same way, but it does include some additional, Code Search-specific information.

For more information, please see

<http://www.google.com/support/webmasters/bin/topic.py?topic=12640>

1.6 Blog Search Ping Introduction

The Google Blog Search Pinging Service is a way to inform Google Blog Search of blog updates. These updates are then published and shared with other search engines to allow them to discover the changes to your blogs. In addition, Google Blog Search will add submitted blogs to the list of blogs it needs to crawl and index.

Google Sitemap Generator allows users who frequently update their blog to automatically inform Google Blog Search about changes to their blogs. Blogging

provider admins can also use it to notify Google of changes to blogs on their platform(s).

For more information, please see

http://www.google.com/help/blogsearch/about_pinging.html

2 Google Sitemap Generator Introduction

2.1 Basic Introduction

Usually, when website masters are to provide sitemap function on their site, the common problems they meet are:

- Hard to construct the sitemap manually, since there are so many pages on the site, and it keeps to grow, new pages are added, old pages are deleted, a few pages' URLs are changed and other pages' URLs are dynamically generated. The sitemap file need to be generated, splitted and updated automatically.
- After sitemap files have been regenerated periodically to pick up new content on your webserver, the search engines should be informed immediately so they can crawl new pages as soon as possible.

To make it easier for webmasters to reap the benefits of sitemaps, we have designed a new tool that automates the Sitemap generation process. The Google Sitemap Generator can generate, maintain and refresh Sitemaps automatically, as well as submit new sitemap files to search engines. The Google Sitemap Generator provides rich configuration options. Webmasters can submit sitemaps to any search engine (that supports sitemaps) by adding the search engine entry.

To guard against accidentally leaking out information on private urls, you can use blacklist patterns to prevent them from being added to a Sitemap. Also, you can set the Sitemap refresh interval, url life time and many more options.

2.2 Mechanism and Limitation

The Google Sitemap Generator collects URLs and their priorities by monitoring the traffic your website receives. It includes two components: a filter plugin and a Sitemap generator.

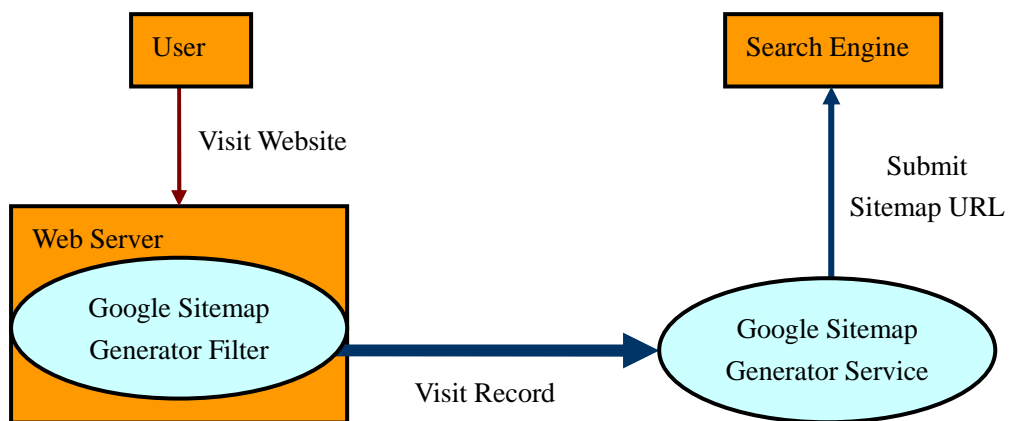


Fig. 1 the architecture of the Google Sitemap Generator

The Filter in your web server can track all visits to your website and send a digest to the Sitemap generator. The sitemap generation service will store all digests effectively and refresh/submit sitemaps regularly. The most work will be done by the sitemap generation service, which runs as a separate process to prevent negatively impact your webserver performance.

Currently, we only developed filter for Internet Information Server 6.0, so Google Sitemap Generator can only be installed on Windows 2003 Server with Internet Information Server 6.0. We will release filter for Apache on Linux in future.

3 Install Google Sitemap Generator (Windows + IIS)

3.1 System Requirement

1. OS:

- Microsoft Windows 2003 Server (English/Chinese, 32b/64b) with Internet Information Service 6.0 installed
- Microsoft Windows 2008 Server (32b/64b) with Internet Information Service 7.0 installed
- Red Hat Enterprise Linux 3 (32b/64b) with Apache 1.3 installed
- Red Hat Enterprise Linux 4 (32b/64b) with Apache 2.0 installed
- SUSE Linux Enterprise Server 10.0 (32b/64b) with Apache 2.2 installed
- Debian etch r0 with Apache 2.2.3 installed
- Fedora 7 with Apache 1.3 installed
- Mandriva 2007 with Apache 2.2 installed
- CentOS 4.6 with Apache 2.0.52 installed

- Ubuntu 6.10 (32b/64b) with Apache 2.2 installed

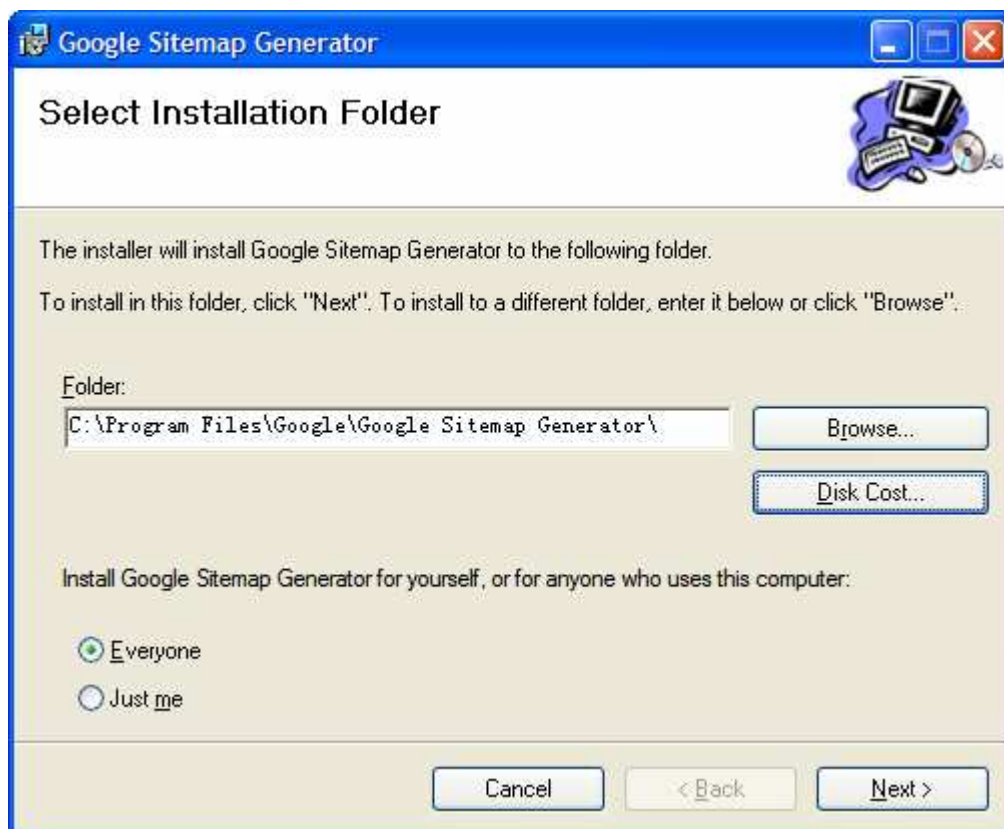
2. Space:

- About 100M to 1G free disk space (depends on the unique URL number on your website).

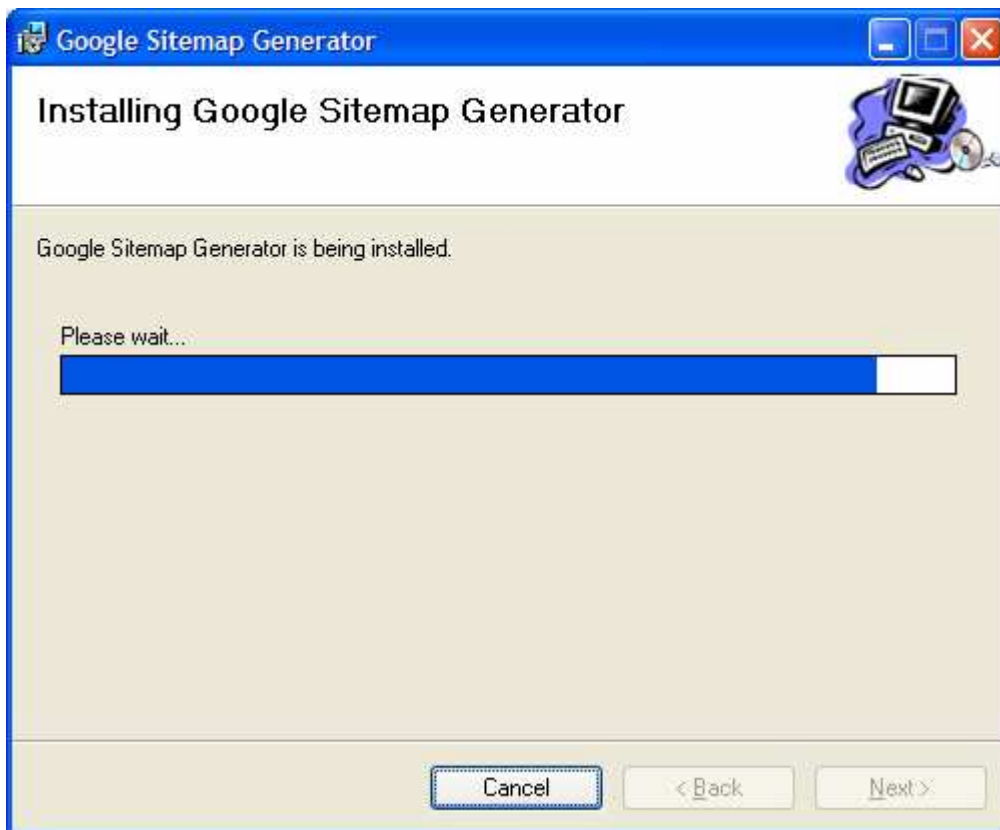
3.2 Installation Steps

Step 1: Run sitemap_setup.msi file to launch Google Sitemap Generator setup program.

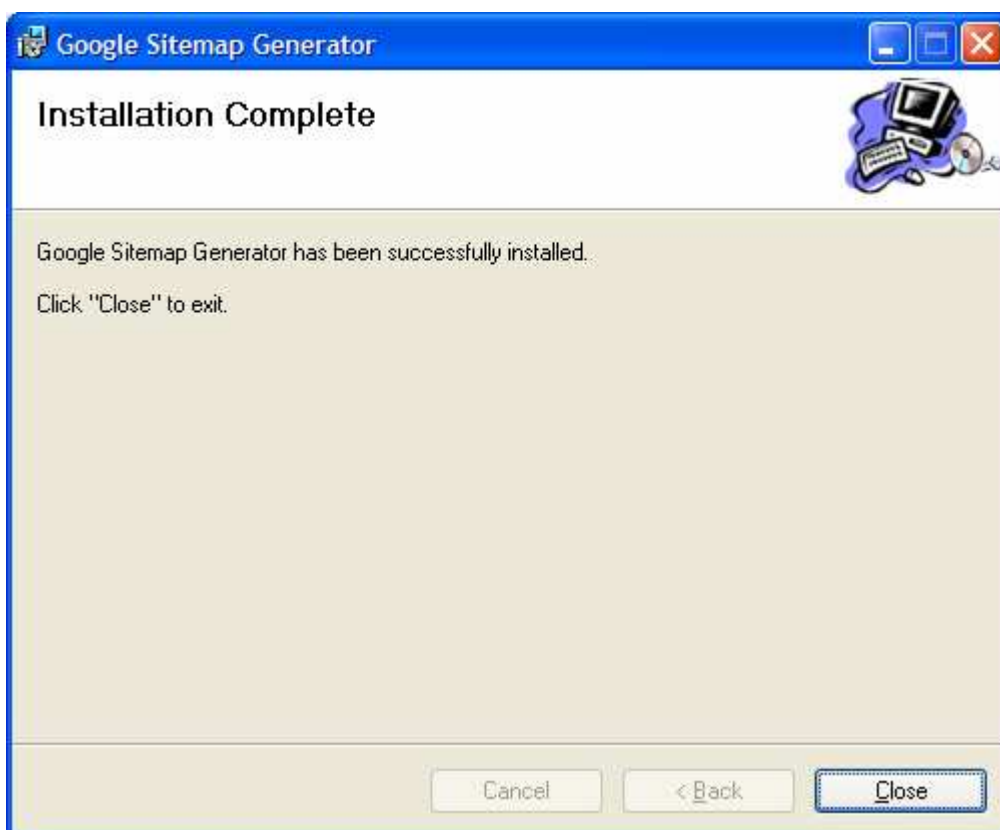
Step 2: Select installation target folder. Google Sitemap Generator needs some disk space to cache historical data. The required disk space depends on the unique URL number of your website. For example, Google Sitemap Generator will require about 1G disk space if your website has about 1M URLs. You can set the max URL number in the web configuration page of Google Sitemap Generator after the installation.



Step 3: Google Sitemap Generator will begin to install required Windows backend service and Internet Information Service Filter. This process will require about tens of seconds.



Step 4: Confirm the installation is finished.

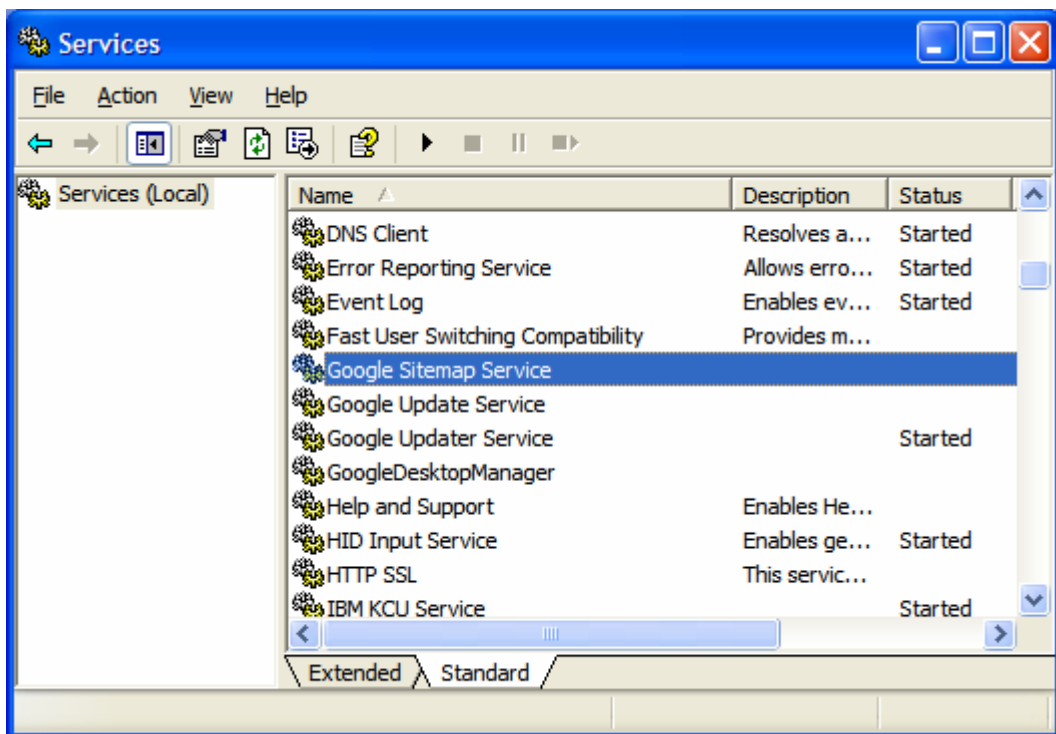


Step 5: Our configuration UI will automatically start up after step 3, you can also

access it from <http://localhost:8181/> . To do the configuration, see [Configuration Part](#). Make sure it has right setting for you before continue.

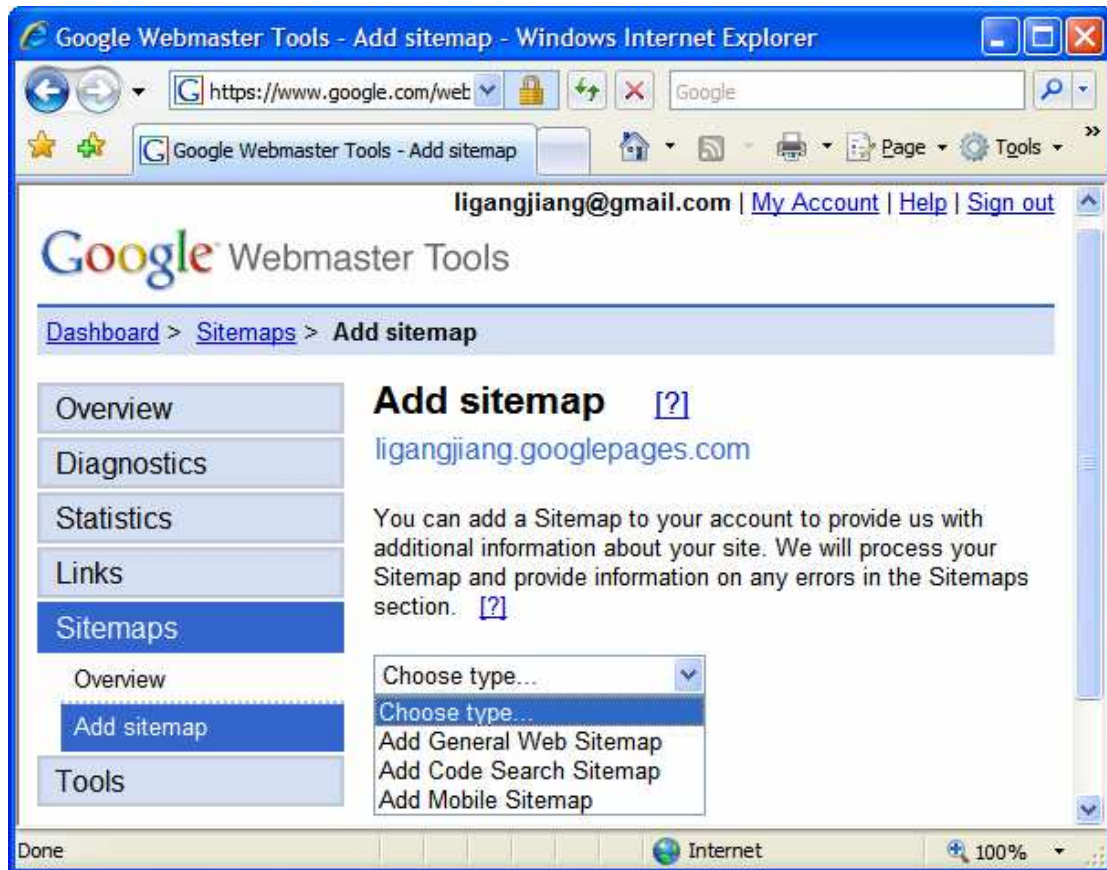
Note: News Sitemap File is disabled by default. If your website is a source of Google News Search, please pickup the “enable” checkbox in the configuration page.

Step 6: Launch Windows Service Management Console from Windows Control Panel. Start “Google Sitemap Service”, and restart “World Wide Web Publishing” Service so filter of Google Sitemap Generator will take effect.



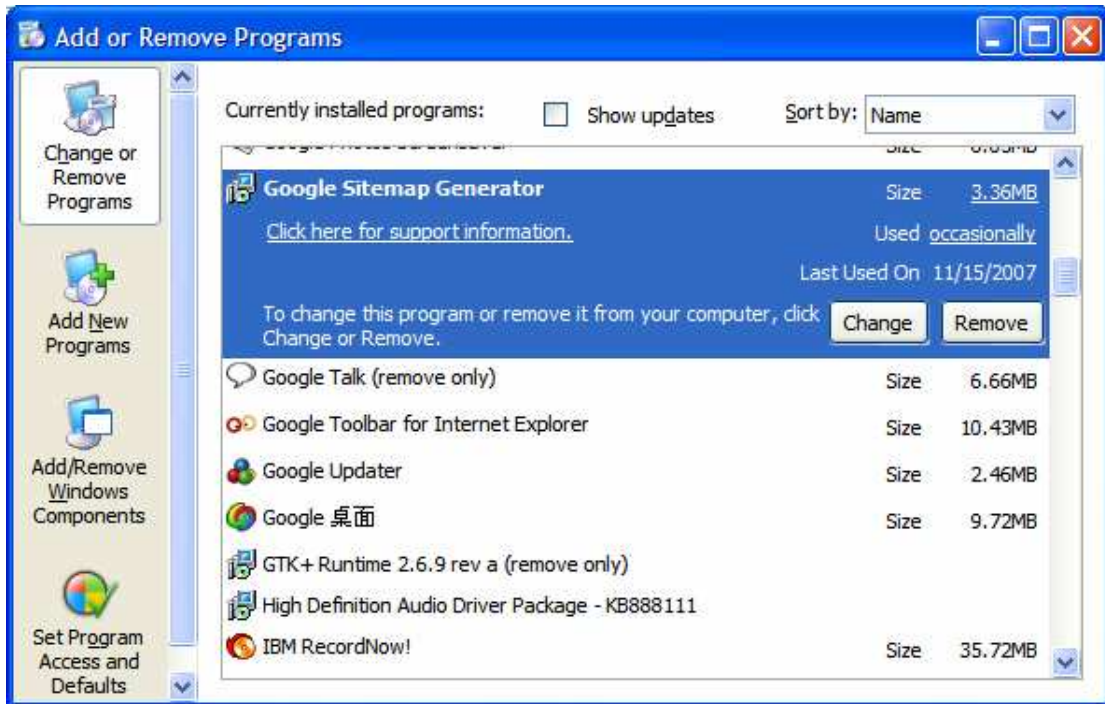
After Google Sitemap Generator Service is started, it will create sitemap files in the root directory of your websites. If you didn't change the default setting, the default file name of these sitemap file will be “sitemap.xml”(for normal website file) or “news_sitemap.xml” (for news sitemap file).

Step 7: Google Sitemap Generator can submit the web(common) sitemap files to search engines automatically. But the news sitemap file has to be manually submitted to Google Search Engine in Google Webmaster Central (<http://www.google.com/webmasters/tools>). And you have to pass the verification in order to make sure you're the administrator of the website.

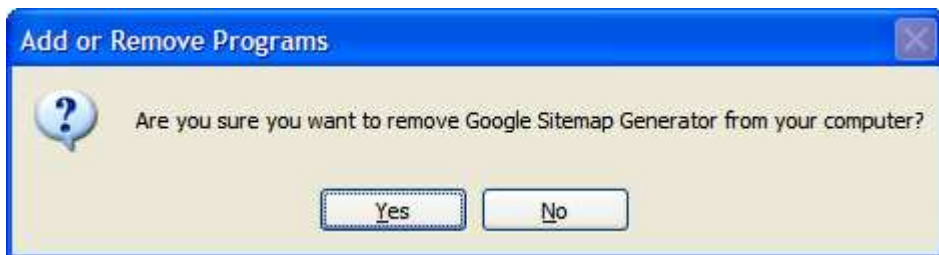


3.3 Uninstallation Steps

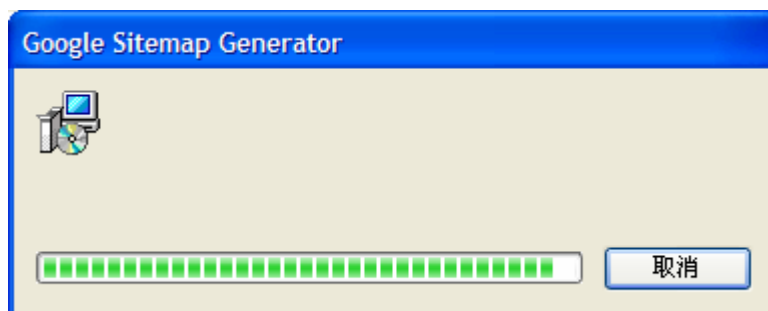
Step 1: Start “Add or Remove Programs” from Windows Control Panel. Find “Google Sitemap Generator” and select “Remove”.



Step 2: Confirm the uninstallation.



Step 3: Uninstallation program will run tens of seconds, including removing filter of “Google Sitemap Generator” and restart “World Wide Web Publishing” Service



Step 4: The historical data and configuration file will not be removed by uninstallation program automatically. Keep these files if you plan to install “Google Sitemap Generator” again, or delete them manually (default path is C:\Program Files\Google\Google Sitemap Generator).

4 Install Google Sitemap Generator (Linux + Apache)

4.1 System Requirement

1. OS:

- Red Hat Enterprise Linux 3 (32b/64b) with Apache 1.3 installed
- Red Hat Enterprise Linux 4 (32b/64b) with Apache 2.0 installed
- SUSE Linux Enterprise Server 10.0 (32b/64b) with Apache 2.2 installed
- Debian etch r0 with Apache 2.2.3 installed
- Fedora 7 with Apache 1.3 installed
- Mandriva 2007 with Apache 2.2 installed
- CentOS 4.6 with Apache 2.0.52 installed
- Ubuntu 6.10 (32b/64b) with Apache 2.2 installed

2. Space:

- About 100M to 1G free disk space (depends on the unique URL number on your website).

4.2 Installation Steps

Step 1: download and extract the install package.

```
sudo tar -zxvf sitemap-install.tar.gz
```

Notes: please extract the package using root account, since the program have to be installed by root.

Step 2: run the install script, which require root permission.

```
sudo sitemap-install/install.sh --apache-bin=[apache binary path]
```

Step 2.1: check the path of Apache.

You can input the path of Apache by using the script command line parameter 'apache-bin'. If not, the install script will automatically find the path of Apache, and require confirm:

```
Which is apache bin?[/usr/sbin/apache2]
```

You can accept the value by directly press Enter, or input the path.

Step 2.2: check the parameters of Apache.

The install script will output some parameters of the Apache program. Please check them and stop the installation immediately if any of them is wrong.

```
*****
From your apache binary, we have detected that,
1) Apache version is: [Version 2.0]
2) Apache architecture is: [32 bits]
```



```
3) Apache root configuration file is: [/etc/apache2/apache2.conf]
```

```
4) Apache pid file is: [/var/run/apache2.pid]
```

If you find any thing above is incorrect, please DO NOT continue.

```
*****
```

```
Do you want to proceed? [N/y]
```

If nothing is wrong, please enter 'y' to continue.

Step 2.3: the install script will display the installation paths of the program:

```
*****
```

This application assumes that,

```
1) Apache httpd is installed, and DSO is enabled.
```

Following directories will be used by this application:

```
1) /var/spool/google-sitemap-generator
```

```
2) /usr/share/google-sitemap-generator
```

```
3) /etc/google-sitemap-generator
```

```
4) /var/lock/google-sitemap-generator
```

Following files will be used by this application:

```
1) /usr/sbin/google-sitemap-generator-ctl
```

```
2) /var/log/google-sitemap-generator.log
```

```
3) /var/run/google-sitemap-generator.pid
```

You could remove all application files by running following command.

```
sudo /usr/share/google-sitemap-generator/uninstall.sh
```

```
*****
```

```
Do you want to proceed? [N/y]
```

You can choose to continue or not.

Step 2.4: the install script will automatically copy the program files and configure the system.

Step 2.5: the installation finish. The configuration UI will start automatically (See [Configure Part](#)).

The most important thing is that if there are multiple Apache servers installed in the system, the install script may choose a wrong one. You have to input the parameters directly, instead of accepting the default value that the install script provides.

Step 3: Some program folders and files will be automatically created in your system:

- 1) `/usr/share/google-sitemap-generator`, which is the path of program files.
- 2) `/var/spool/google-sitemap-generator`, which is the path of internal database for URLs storage. In the running process, it will occupy about 100M to 1G disk space according to the number of URLs.

- 3) `/etc/google-sitemap-generator/sitesettings.xml`, which is the configuration file.
- 4) `/var/log/google-sitemap-generator.log`, which is the log file.
- 5) `/usr/sbin/google-sitemap-generator-ctl`, which is the control script of the program.
- 6) `/etc/init.d/google-sitemap-generator`, which is the script used by system service. This script can automatically start the program when system starts.
- 7) Some other files that may be used:
`/var/lock/google-sitemap-generator`
`/var/run/google-sitemap-generator.pid`.

Step 4: the Sitemap filter will be added to the Apache server.

The install script will add a new line to the configuration file of Apache so that Apache can load the Sitemap filter when it runs:

```
LoadModule google_sitemap_genrator_module
/usr/share/google-sitemap-generator/mod_sitemap.so
```

It will backup the original configuration file to `/usr/share/google-sitemap-generator/httpd.install.conf`.

The modification will be automatically rollback when the Generator is uninstalled. The uninstall script will backup the modified configuration file to `/usr/share/google-sitemap-generator/httpd.uninstall.conf`.

Step 5: Configure Google Sitemap Generator.

After the installation, the configuration web page will be opened in your browser. You can access the configuration page at any time through the address <http://localhost:8181>. For more information, please see [Configure Part](#).

The Advance user can manually edit the configuration file `/etc/google-sitemap-generator/sitesettings.xml`.

Notes: most configuration changes require restart of the Generator to take effect.

Step 6 : You can control the Generator service by running `google-sitemap-generator-ctl`. It's under `/usr/sbin`.

This script requires root permission. For example, you can restart the service by running:

```
sudo google-sitemap-generator-ctl -restart
```

Besides the 'restart' parameter, you can use 'start' and 'stop' to start or stop the service.

Notes: the Generator service will be automatically started when the system is started. And `google-sitemap-generator-ctl` cannot control the sitemap filter in Apache, you have to use the Apache controller to do it.

4.3 Uninstallation Steps

Step 1: run the uninstall script (need root permission)

```
sudo /usr/share/google-sitemap-generator/uninstall.sh
```

During the uninstall process, you may be asked to stop the Apache server since the sitemap filter cannot be removed when the Apache server is running.

Step 2: By default, all the generated data and configuration file will not be removed by the uninstall script. If you don't want to install the new version of Google Sitemap Generator in future, you can remove these files by choosing 'y' at the last step of the Uninstallation.

5 Configure Google Sitemap Generator

5.1 Introduction

The configuration page(we call it Site Setting Editor) can be accessed from [http://localhost:\[settingport\]/](http://localhost:[settingport]/) once the server is started; you can also directly modify the SiteSettings.xml file in the application install directory (default to C:\Program Files\Google\Google Sitemap Generator).

The setting port is default to 8181 (you can check the sitesettings.xml or server log file if you can't find the correct port), you can configure it in the setting page too.

The following is the login page you can see when open the URL:



Site Setting Editor

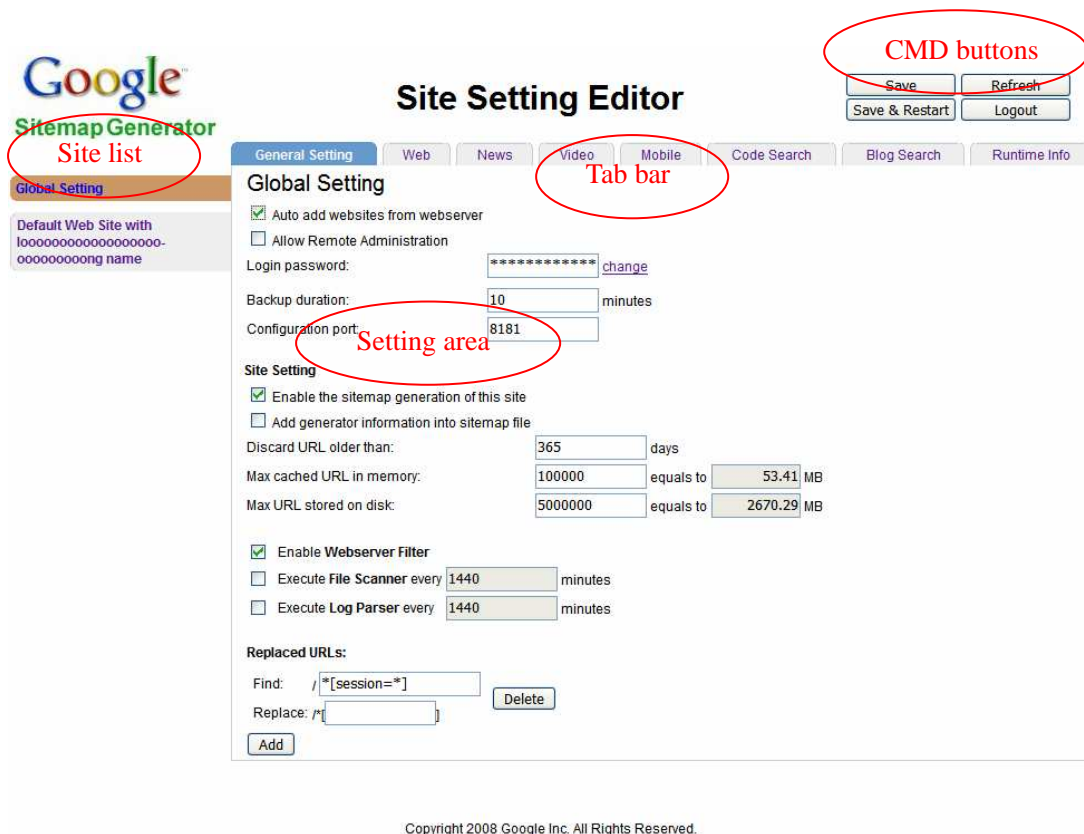
Username:

Password:

Copyright 2008 Google Inc. All Rights Reserved.

You have to input the password to login the setting page. The initial password is 'admin', and you can change the password in the setting page after login succeeds.

This is the General Global setting page you will see when first login:



There are four parts on the page:

➤ Site list

On the left side, you can choose the site setting which you want to edit. All the sites you configured in the web server (IIS or Apache) will be listed here. The first one is for the global setting, which provides a way to set multiple sites in one place. Anything that you changed here will be auto set to any of the other site settings that are set to inherit from it.

➤ Tab bar

On the top side, you can choose which Tab you want to watch/edit. The first Tab is for general setting of a site, and the last Tab is to display runtime information from server. All the other Tabs are for sitemaps. Up to now, we support five types of sitemaps: Web, News, Video, Mobile, Code Search, and one type of ping (another way to inform search engine that the site has new content): Blog Search. We may add more sitemaps in future.

➤ CMD button

On the top right corner, there are four action buttons. The actions you can do includes:

- Save: save the changed settings to server.
- Refresh: get the newest settings from server; It will discard all the unsaved settings, so be careful.

- Save & Restart: save the settings to server, and then restart the server, after the server restarts, this page will be redirected to the right URL automatically no matter whether the setting port has been changed.
- Logout: the current session will be logout, and the page will be redirect to login page.

Note: here 'server' means the simple web server that is embedded in the Google Sitemap Generator, to provide service for web-based configuration.

➤ Setting area

This area is for the site/sitemap setting.

There are three types of settings on the configuration pages:

1. Checkbox setting:

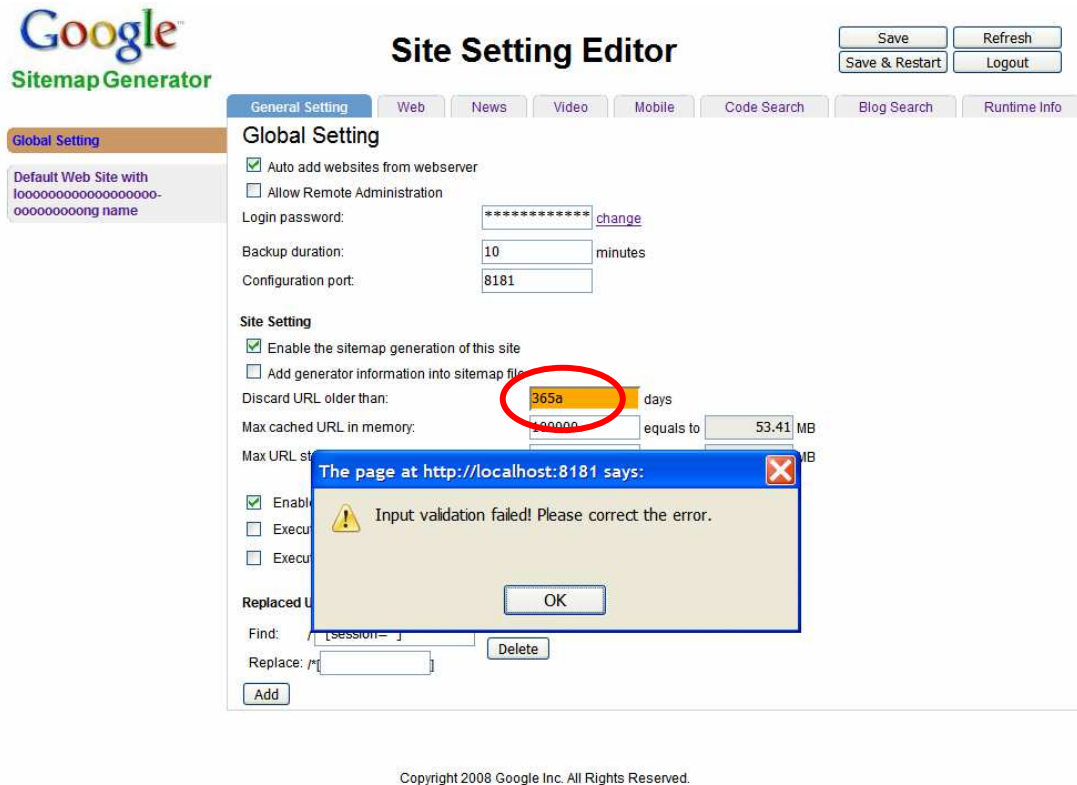
☒ Auto add

This type of setting is the simplest; 'checked' means set the item to 'true', 'unchecked' means set to 'false'.

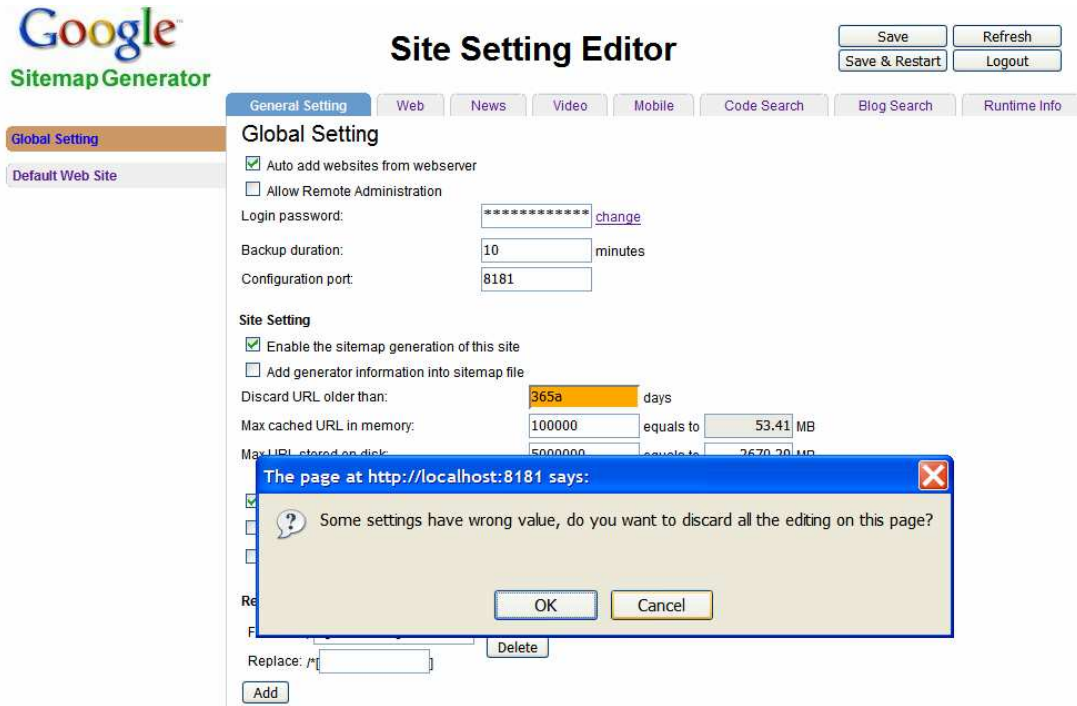
2. Text setting:

Setting port:

This type of setting is for single text, number, and date items. When user edit the value, it will be verified first. If the value is invalid, the input field will be highlight and require correcting. If you click the 'save' button without correcting the error input, the same warning dialog will popup again and the 'save' action will be ignored.



If you want to switch to another tab page or another site when there is invalid setting on current Tab, another dialog will popup (see below). You can choose to discard the wrong editing (actually, all the editing on the page will be discarded), or stay at this page.



Copyright 2008 Google Inc. All Rights Reserved.

Some items have additional grey box to provide reference value for your setting. For example, if you change the value for 'Max URL stored on disk' item, the grey box on the right side will show the max space that may be occupied by URLs. This will help you to choose a suitable value for the item.

Max URL stored on disk: 5000000 equals to 2670.29 MB

3. List setting:

Replaced URLs:

Find: /[*[session=*][id]
 Replace: /*[] [] Delete

This type is for group of variable number of items. You can click the 'add' button to add a new line, or click the 'delete' button on the right to delete this line. The input in each line will also be verified. The following are some other examples.

Included URLs:**Excluded URLs:**

URL Pattern: /	robots.txt	<input type="button" value="Delete"/>
URL Pattern: /	*.jpg	<input type="button" value="Delete"/>
URL Pattern: /	*.gif	<input type="button" value="Delete"/>
URL Pattern: /	*.png	<input type="button" value="Delete"/>
URL Pattern: /	*.css	<input type="button" value="Delete"/>
URL Pattern: /	*.js	<input type="button" value="Delete"/>
URL Pattern: /	*.swf	<input type="button" value="Delete"/>
URL Pattern: /	*?*	<input type="button" value="Delete"/>
URL Pattern: /	*password*	<input type="button" value="Delete"/>

Notify following search engine URLs:

URL: http://	www.google.com/webmasters/sitemaps/ping?sitemap=	<input type="button" value="Delete"/>
URL: http://	search.yahooapis.com/SiteExplorerService/V1/ping?sitemap=	<input type="button" value="Delete"/>
URL: http://	submissions.ask.com/ping?sitemap=	<input type="button" value="Delete"/>
URL: http://	api.moreover.com/ping?u=	<input type="button" value="Delete"/>
URL: http://	webmaster.live.com/ping.aspx?siteMap=	<input type="button" value="Delete"/>

Note: all the setting items will have a default value after first installation. Besides, the uninstall script will not remove the setting configuration XML file as well as the generated sitemap files, and the setting file will be recognized by install script. So when you do the next new installation, you will not need to reconfigure the setting.

5.2 Site setting describe

5.2.1 Global Site Setting:

The screenshot shows the 'Site Setting Editor' for Google Sitemap Generator. It has a sidebar with 'Global Setting' and 'Default Web Site' options. The main panel is titled 'Global Setting' and contains several sections: 'General Setting' with checkboxes for 'Auto add websites from webserver' (checked) and 'Allow Remote Administration' (unchecked), a 'Login password' field with a 'change' link, and input fields for 'Backup duration' (10 minutes) and 'Configuration port' (8181). The 'Site Setting' section includes checkboxes for 'Enable the sitemap generation of this site' (checked) and 'Add generator information into sitemap file' (checked), followed by 'Discard URL older than' (365 days), 'Max cached URL in memory' (100000, equals to 53.41 MB), and 'Max URL stored on disk' (5000000, equals to 2670.29 MB). There are also checkboxes for 'Enable Webservice Filter' (checked), 'Execute File Scanner every' (1440 minutes), and 'Execute Log Parser every' (1440 minutes). The 'Replaced URLs' section has 'Find' and 'Replace' fields with 'Add' and 'Delete' buttons. At the top right are 'Save', 'Refresh', 'Save & Restart', and 'Logout' buttons.

Copyright 2008 Google Inc. All Rights Reserved.

Auto add websites from webserver: let Google Sitemap Generator automatically add all the sites that are in the IIS or Apache.

Allow Remote Administration: allow administrator can login from remote computer to configure the site setting.

Login password: it's a special setting component. If you want to change the value, you cannot directly input the login password in the text box, instead, you have to click the 'change' link, input old password once, new password twice, and click the 'save' button (see below). The password length is limited from 6 to 50. Once you click the 'save' button (or press 'enter' key in the last input box), the change will be immediately submitted to the configuration server.

The screenshot shows a password change dialog box with three input fields: 'Input old password:', 'Input new password', and 'Input again'. Below the fields are 'Save' and 'Cancel' buttons.

Backup duration: how long to backup the sitemap URLs from memory to disk. It must be in range [10, 2000000) minutes.

Notes: It's the backup for URLs that cached in the memory, in case to avoid suddenly shutdown. When doing backup, all the memory space will be written to disk, but the memory itself will not be clean.

Configuration port: which port the Google Sitemap Generator is listening for the configuration. It must be in range (0, 65536).

Please refer to [4.2.2 Normal Site Setting](#) for the other settings.

5.2.2 Normal Site Setting:

The screenshot shows the 'Site Setting Editor' interface for the Google Sitemap Generator. The interface includes a sidebar with 'Global Setting' and 'Default Web Site' options. The main panel is titled 'Site Setting Editor' and contains several tabs: 'General Setting', 'Web', 'News', 'Video', 'Mobile', 'Code Search', 'Blog Search', and 'Runtime Info'. The 'General Setting' tab is active, showing the 'Site Setting' section. This section includes a checkbox for 'Inherited from "Global Setting"', which is checked. Below this, there are input fields for 'Host Name' and 'Log Path'. The 'Site Setting' section also includes a checkbox for 'Enable the sitemap generation of this site', which is checked, and a checkbox for 'Add generator information into sitemap file', which is also checked. There are three input fields for 'Discard URL older than:', 'Max cached URL in memory:', and 'Max URL stored on disk:', each with a corresponding 'days' or 'MB' unit. The 'Max cached URL in memory:' field is set to 100000, which equals 53.41 MB. The 'Max URL stored on disk:' field is set to 5000000, which equals 2670.29 MB. There are also checkboxes for 'Enable Webserver Filter', 'Execute File Scanner every 1440 minutes', and 'Execute Log Parser every 1440 minutes'. At the bottom, there is a 'Replaced URLs' section with 'Find' and 'Replace' input fields, a 'Delete' button, and an 'Add' button. The 'Find' field contains the text '/*[session=*]' and the 'Replace' field is empty.

Copyright 2008 Google Inc. All Rights Reserved.

Inherited from “Global setting”: If the setting is true, all the settings (except some site special setting) on this page are synchronized with global setting. The setting will on each sitemap setting page, too.

Note: site special setting refers to some setting that is always different for each site, such as “Host Name” and “Log Path”

Host Name: the host name of this site.

Log Path: the Apache or IIS log path of this site.

Enable the sitemap generation of this site: If true, Google Sitemap Generator will serve the site.

Discard URL older than: if the page's content of this URL is older than this setting, it will not be included in the sitemap files. It must be in range (0, 200000000) days.

Max cached URL in memory: the max URL numbers allowed to cache in the memory. The bigger value will apply better performance for Google Sitemap Generator, while more memory space will be consumed. It must be in range (0, 200000000).

Max URL stored on disk: the max URL numbers allowed to store in sitemap files. The suitable value depends on how large your site is and how much disk space you can provide for storing the URLs. It must be in range (0, 200000000).

Note: These two space settings only limit the size of internal URL database, not include the space that occupied by generated sitemaps and the log file of the Generator. When the space of the database grows to the max limitation, older URLs will be discarded in order to store new URLs. Don't worry! It will not remove the discarded URLs from the search engine, since Google only use sitemap to find new URLs (it will use another way to judge if the URL has been invalid).

Replaced URLs: for some URL, you may want to store different value to the sitemap file other than the original value that capture from IIS or Apache. This will help the webmaster to hide some privacy information from the search engines, and also help Google Sitemap Generator to reduce the URLs that refer to the same page (especially for the dynamic generated web pages).

Find: it's the pattern to match the original URL value, refer to [4.3.5](#) for the pattern rule.

Replace: it's the replace value for the content that matches the 'find' pattern.

5.2.3 Web Sitemap Setting:

Google Sitemap Generator

Site Setting Editor

Save Refresh
Save & Restart Logout

General Setting Web News Video Mobile Code Search Blog Search Runtime Info

Global Setting
Default Web Site

Web Sitemap Setting

☒ Enable this sitemap generation
☒ Compress sitemap file
☐ Include sitemap URL in robots.txt

Sitemap file name:

Update sitemap file from: (example: 2008-01-01 12:34:56)

Update sitemap file every: minutes

Max URL number per sitemap file:

Max file size per sitemap file: KB

Included URLs:

Excluded URLs:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

Notify following search engine URLs:

URL:

URL:

URL:

URL:

URL:

Copyright 2008 Google Inc. All Rights Reserved.

Enable this sitemap generation: if true, the web sitemap for this site is enabled.

Compress sitemap file: if true, the sitemap files will be compressed. Up-to-now, we use ZLIB to do the job.

Sitemap file name: the generated sitemap file name. If more than one sitemap file are generated, it will be the index file name (only news sitemap is limited to one file, the max limitation for sitemap files of other type is 1000). The valid input includes any ASCII word character ([a-zA-Z0-9_]), with '.xml' as tail.

Update sitemap file from: the date is the start time of the web sitemap service. The valid format is "yyyy-mm-dd hh:mm:ss".

Update sitemap file every: it defines how long the sitemap file will be regenerated. It must be in range [10, 2000000) minutes.

Max URL number per sitemap file: it limits each sitemap file's size. It must be in range [1, 50000].

Max file size per sitemap file: another way to set the file's size. It must be in range (0, 10485760].

Note: these two items has the same functionality, the actual size will be the minimum one of these two values. Besides, the file size is for the size before compress, the reason is that search engines will check the size after uncompressed. So if you choose to compress the files, they will occupy much smaller space on the disk.

Included URLs: all the URLs that match these patterns will be included in the sitemap files, all the other URLs will be ignore.

URL Pattern: the pattern of the URL, refer to [4.3.5](#) for pattern rule

Excluded URLs: all the URLs that match these patterns will be excluded from the sitemap files, all the other URLs will be included

URL Pattern: the pattern of the URL, refer to [4.3.5](#) for pattern rule

Note: If neither of Included URLs and Excluded URLs is set, all the URLs will be included. If both are set, the URLs that match both rules will be excluded, and the URLs that match neither will be excluded too.


The “Included” and “Excluded” setting also affect the sitemap size directly. If both are set to empty, all the URLs in the internal database will be put into the sitemap. Be careful to set the right rules before enabling a sitemap.

Notify following search engine URLs: the search engines' URL to accept sitemap.

URL: the http URL

Note: https protocol is not supported.

5.2.4 News Sitemap Setting:



Site Setting Editor

Save

Refresh

Save & Restart

Logout

General Setting

Web

News

Video

Mobile

Code Search

Blog Search

Runtime Info

Global Setting

Default Web Site

News Sitemap Setting

☐ Enable this sitemap generation

☐ Compress sitemap file

Sitemap file name:

Update sitemap file from: (example: 2008-01-01 12:34:56)

Exclude news older than: minutes

Update sitemap file every: minutes

Max URL number per sitemap file:

Max file size per sitemap file: KB

Included URLs:

URL Pattern: /

Delete

URL Pattern: /

Delete

Add

Excluded URLs:

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

URL Pattern: /

Delete

Add


Copyright 2008 Google Inc. All Rights Reserved.

Exclude news older than: since news page need stricter time limit, this setting has the same effect as ‘Discard URL older than’ but has more precise time unit. The news page URLs will be discarded if it’s older than either of these two settings. It must be in range (0, 200000000).

Please refer to [4.2.3 Web Sitemap Setting](#) for the other settings.

31

5.2.5 Video Sitemap Setting:



Site Setting Editor

Save

Refresh

Save & Restart

Logout

General Setting

Web

News

Video

Mobile

Code Search

Blog Search

Runtime Info

Global Setting

Default Web Site

Video Sitemap Setting

☐ Enable this sitemap generation

☒ Compress sitemap file

Sitemap file name:

Update sitemap file from: (example: 2008-01-01 12:34:56)

Update sitemap file every: minutes

Max URL number per sitemap file:

Max file size per sitemap file: KB

Included URLs:

URL Pattern:

Delete

Add

Excluded URLs:

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:


Delete

Add

Copyright 2008 Google Inc. All Rights Reserved.

Please refer to [4.2.3 Web Sitemap Setting](#) for the other settings.

5.2.6 Mobile Sitemap Setting:



Site Setting Editor

Save

Refresh

Save & Restart

Logout

General Setting

Web

News

Video

Mobile

Code Search

Blog Search

Runtime Info

Global Setting

Default Web Site

Mobile Sitemap Setting

☐ Enable this sitemap generation

☒ Compress sitemap file

Sitemap file name:

Update sitemap file from: (example: 2008-01-01 12:34:56)

Update sitemap file every: minutes

Max URL number per sitemap file:

Max file size per sitemap file: KB

Included URLs:

Add

Excluded URLs:

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:


Delete

Add

Copyright 2008 Google Inc. All Rights Reserved.

Please refer to [4.2.3 Web Sitemap Setting](#) for the other settings.

5.2.7 Code Search Sitemap Setting:



Site Setting Editor

Save

Refresh

Save & Restart

Logout

General Setting

Web

News

Video

Mobile

Code Search

Blog Search

Runtime Info

Global Setting

Default Web Site

Code Search Sitemap Setting

☐ Enable this sitemap generation

☒ Compress sitemap file

Sitemap file name:

Update sitemap file from: (example: 2008-01-01 12:34:56)

Update sitemap file every: minutes

Max URL number per sitemap file:

Max file size per sitemap file: KB

Included URLs:

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

URL Pattern:

Delete

Add

Excluded URLs:

Add

Copyright 2008 Google Inc. All Rights Reserved.

Please refer to [4.2.3 Web Sitemap Setting](#) for the other settings.

34

5.2.8 Blog Search Ping Setting



The screenshot shows the 'Site Setting Editor' interface for the Google Sitemap Generator. The 'Blog Search' tab is selected, displaying the 'Blog Search Ping Setting' configuration. The interface includes a sidebar with 'Global Setting' and 'Default Web Site' options. The main content area contains a checkbox for 'Enable this sitemap generation', a text input for 'Execute Ping every' (set to 30 minutes), and sections for 'Included URLs' and 'Excluded URLs', each with an 'Add' button. At the top right, there are buttons for 'Save', 'Refresh', 'Save & Restart', and 'Logout'. The footer indicates 'Copyright 2008 Google Inc. All Rights Reserved.'

Google Sitemap Generator

Site Setting Editor

Save Refresh
Save & Restart Logout

General Setting Web News Video Mobile Code Search **Blog Search** Runtime Info

Global Setting
Default Web Site

Blog Search Ping Setting

☐ Enable this sitemap generation

Execute Ping every 30 minutes

Included URLs:

Excluded URLs:

Copyright 2008 Google Inc. All Rights Reserved.

Execute Ping every: It must be in range [10, 2000000) minutes.

Please refer to [4.2.3 Web Sitemap Setting](#) for the other settings.

5.2.9 Runtime Info:

For Global setting, it will display Application level runtime information.



Copyright 2008 Google Inc. All Rights Reserved.

We will show the memory and disk space occupied by the application (not include the sitemaps and the log file) and the start time of it. You can click the 'Refresh Runtime Info' button the get the latest runtime information (it will refresh all the runtime info pages, the 'Refresh' button has the same effect, but it will refresh the setting Tabs, too).

For sites, it will display the runtime information of the services of the site.

The screenshot shows the 'Site Setting Editor' interface for Google SitemapGenerator. The 'Runtime Info' tab is selected, displaying a table of runtime information for various services. The table is organized into sections for different services, each with a 'Last Update URLs Count', 'Last Running Result', 'Last Running Time', and 'Last Update URLs Count'.

Service	Last Update URLs Count	Last Running Result	Last Running Time	Last Update URLs Count
Runtime Info for site	0			
URLs in database:	0			
URLs in tempfile:	0			
URLs in memory:	0			
Host Name:	Undetermined			
Memory Used For All Services:	0 bytes			
Disk Used For All Services:	0 bytes			
Webserver Filter Service:				
Last Update URLs Count:	0			
Blog Search Ping Service:				
Last Running Result:	failed			
Last Running Time:	-1			
Last Update URLs Count:	N/A			
Log Parser Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			
File Scanner Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			
Web Sitemap Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			
News Sitemap Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			
Video Sitemap Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			
Mobile Sitemap Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			
Code Search Sitemap Service:				
Last Running Result:	N/A			
Last Running Time:	N/A			
Last Update URLs Count:	0			

Copyright 2008 Google Inc. All Rights Reserved.

Currently, we will show the site services summary information such as the total URLs count of all services generated the most used host name of the site, and the memory/disk space for all the services. We will also show the information of each service, such as running result, the last running time, and the count of URLs that generated by the service.

Note: URLs in tempfile are the URLs that move from memory to disk but haven't been merged into the database. It's not the same as the backup for the memory, the memory will be clean when URLs are written to the tempfile.

If the site is not running, we will give the warning (see below).

Site Setting Editor

Save Refresh
Save & Restart Logout

Global Setting

Default Web Site

General Setting

Web

News

Video

Mobile

Code Search

Blog Search

Runtime Info

Runtime Info for site

Site services are not running

Refresh
Runtime Info

URLs in database:		
URLs in tempfile:		
URLs in memory:		
Host Name:		
Memory Used For All Services:		bytes
Disk Used For All Services:		bytes
Webserver Filter Service:		
Last Update URLs Count:		
Blog Search Ping Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
Log Parser Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
File Scanner Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
Web Sitemap Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
News Sitemap Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
Video Sitemap Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
Mobile Sitemap Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		
Code Search Sitemap Service:		
Last Running Result:		
Last Running Time:		
Last Update URLs Count:		

5.3 FAQ

5.3.1 How to set the setting port?

Google Sitemap Generator

Site Setting Editor

Save Refresh
Save & Restart Logout

General Setting Web News Video Mobile Code Search Blog Search Runtime Info

Global Setting

Default Web Site with
ooooooooooooooooooooo-
ooooooooooooong name

☒ Auto add websites from webserver
☐ Allow Remote Administration
Login password: ***** [change](#)
Backup duration: 10 minutes
Configuration port: 8181

Site Setting

☒ Enable the sitemap generation of this site
☐ Add generator information into sitemap file
Discard URL older than: 365 days
Max cached URL in memory: 100000 equals to 53.41 MB
Max URL stored on disk: 5000000 equals to 2670.29 MB

☒ Enable Webserver Filter
☐ Execute File Scanner every 1440 minutes
☐ Execute Log Parser every 1440 minutes

Replaced URLs:
Find: /*[session=]*
Replace: /*
[Delete](#)
[Add](#)

Copyright 2008 Google Inc. All Rights Reserved.

5.3.2 How to enable the Web and other sitemaps?

Web Sitemap Setting

☐ Enable this sitemap generation

☐ Compress sitemap file

☐ Include sitemap URL in robots.txt

Sitemap file name:

Update sitemap file from: (example: 2008-01-01 12:34:56)

Update sitemap file every: minutes

Max URL number per sitemap file:

Max file size per sitemap file: KB

Included URLs:

Excluded URLs:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

URL Pattern:

Notify following search engine URLs:

URL:

URL:

URL:

URL:

URL:

5.3.3 How to limit the disk&mem space using?

1. Space that used by internal DB per site

Google Sitemap Generator

Site Setting Editor

Global Setting

☒ Auto add websites from webserver

☐ Allow Remote Administration

Login password: ***** [change](#)

Backup duration: 10 minutes

Configuration port: 8181

Site Setting

☒ Enable the sitemap generation of this site

☐ Add generator information into sitemap file

Discard URL older than: 365 days

Max cached URL in memory: 100000 equals to 53.41 MB

Max URL stored on disk: 5000000 equals to 2670.29 MB

☒ Enable Webserver Filter

☐ Execute File Scanner every 1440 minutes

☐ Execute Log Parser every 1440 minutes

Replaced URLs:

Find: /*[session=*

Replace: /*

Copyright 2008 Google Inc. All Rights Reserved.

2. Space that used by sitemap

5.3.4 I don't know what's the valid value to input, what can I do?

You can refer to [4.2 Site setting describe](#) for detail information. We also add hints on the items, you will see some help information when move the mouse on to the item.

5.3.5 What are the URL pattern language rules?

Pattern	Matched URL example	Describe
/*example*	http://www.example.com http://bbs.example.org	<ul style="list-style-type: none"> This is for general URL pattern field. Start with '/' Mark '*' is the only supported RegExp rule, it matches zero or more characters; Don't add 'http:/' in front of the URL, it will be added to

		rule by default.
/abc[12*]xyz[45]	http://abc123xyz45	<ul style="list-style-type: none"> • Here is some special rules for the pattern of 'find' field in 'URLReplacements' • Using '[']' to mark the part that need to be replaced. • '*' is the only RegExp mark that can used in and out '[']'.
[234][543]	http://abc234xyz543	<ul style="list-style-type: none"> • This is for 'replace' field in 'URLReplacements'. • For each '[']' in 'find' field, a corresponding '[']' part must be occur in 'replace' field, the content in it will be used to replace the part of the URL that match the pattern in the 'find' field.

6 Trouble Shooting

7 Contact

If you have any question or suggestion that you want us to know, please go to the Google Groups:
<http://groups.google.com/group/google-sitemap-generator>