Google Search Appliance – Feeds Protocol

Feature Snippet

SPECIFICATIONS

Related Technologies:

XML – Extensible Markup Language

HTTP – Hypertext Transfer Protocol

Google Search Appliance

Hardware: v4.0 or greater
Software: v4.2 or greater

ORDERING INFORMATION

www.google.com/enterprise

Email appliance1@google.com

Business Overview

The new Google Search Appliance Feeds Protocol enables non-web accessible content to be pushed into the Google Search Appliance with a simple XML conversion. This removes information barriers within corporations, and makes all enterprise content easily accessible and searchable by users. With the feeds protocol, enterprise administrators can push non-web accessible content from sources such as document management, enterprise applications, or legacy systems directly into the Google Search Appliance for indexing and search.

Exporting Content

To feed enterprise content to the Google Search Appliance, it must first be exported from its system of record and converted into the proper format for transport to the Google Search Appliance. The Google Search Appliance Feeds Protocol uses the industry standard Extensible Markup Language (XML) to make generating a feed simple and straightforward.

The Google Search Appliance Feeds Protocol allows for two types of data feeds: content feeds, and URI only feeds. Each document is tracked by its URI as a unique key, and additional parameters are set as attributes in the XML feed file.

Content Feeds

Content feeds allow administrators to push document content in its native form (e.g. HTML, text, Microsoft Word, PDF, Microsoft Powerpoint, etc.) into the Google Search Appliance. This allows non-web accessible documents to be indexed by the search appliance. Text-based content such as HTML and text documents are inserted directly in the `<content>` tag in the body of the XML feed file. Binary documents such as Word, Excel, and PDF files are base64 encoded and the encoded text is inserted in the `<content>` tag in the body of the XML feed file. The file type is specified in the `mimetype` attribute of the `<record>` element.

URI Only Feeds

URI only feeds allow administrators to send a list of http-accessible URI's to the Google Search Appliance Web Crawler, to be crawled and indexed. This allows administrators to give the search appliance URI's that are difficult to discover automatically, but can be crawled by the standard web crawler. The URI's are added to the top of the web crawler queue, and the pages will be crawled, indexed, and managed by the web crawler ongoing.

Feeding Content

Once the desired content feed is generated, the XML feed file is pushed to the Google Search Appliance using the HTTP-POST method. Each unique content repository is referred to as a data source, and can be given a convenient name for reporting. The Feeds Protocol supports both full and incremental feeds. The POST method call consists of the data source name, feed type, and the XML file path.

Example:

```
http://<search_appliance>:19900/xmlfeed/datasource=&
feedtype=&data=<xmluri>
```

Full Feed

A full feed is a push of the complete data source. That is, every time there is a push of a full feed into the appliance, all documents in the data source are pushed. The appliance determines which documents are no longer part of full feed and deletes them from the index.

Example:

A customer creates a full feed and pushes it to the appliance.

- The full feed contains documents D0, D1 and D2. The system serves D0, D1, and D2 along with other crawled documents.
- Later, the administrator creates another full feed containing documents D0, an updated D1, and a new D3 and pushes this feed into the appliance.
- When the feed processing is complete, the system serves D0, the updated D1, and the new D3 along with other crawled documents.
- Document D2 was not part of the second full feed and is removed from the index.

Incremental Feed

An incremental feed deletes documents as it pushes new documents through the feed. Incremental feeds are more efficient for the system and can also help overcome the constraints of a very large data source. A typical use of incremental feeds is to push a full feed initially and incremental feeds for ongoing updates. At any time, a full feed can be pushed to 'reset' the feed content.

Example:

- Day 1 - Administrator pushes a full feed with documents D0, D1 and D2. The system serves D0, D1 and D2.
- Day 2 - Administrator pushes an incremental feed containing an "add" for D3, an "add" for an updated D1 and a "delete" for D2. The system serves D0, updated D1, and D3.
- Day 7 - Administrator pushes a full feed containing documents D0, D7, and D10. The system serves D0, D7, and D10 when the full feed processing is complete.

Serving Results:

The documents that are received via the Google Search Appliance Feeds Protocol are indexed by the search appliance. When a user conducts a search, results including fed documents as well as content from the database and web crawler are displayed to the user. If the fed documents have a web-accessible URI, the user can select the URI and link to the original source document. If the fed document is not web accessible or available through a browser, the user can view the cached document or use the document URI to locate the source document with another means.