

The family of Google Crawlers

Crawlers are the lifeline of Google ranking process. We all know that it acts as a junction for any new or updated content on the web. The pages garnered by the crawlers find their way to the index server. **GOOGLEBOT** is the most recognized as well as the most ubiquitous of all the crawlers on the web.

The intent of this document is to shed light on other members of the Google Crawler family - who, although less popular, but are equally indispensable for the internet search, advertising & marketing. Here is the family composition.

GOOGLE CRAWLERS

GOOGLEBOT – Collects web pages

FRESSHBOT – Collects updated data more frequently

DEEPBOT – Follows every link that it finds

IMAGEBOT – Crawling for the Image search

MEDIABOT – Used to Analyze Adsense pages

ADSBOT – Crawling Adwords landing pages for quality

GOOGLEBOT-MOBILE – Crawling for Mobile pages

GSA-CRAWLER – Used by Google Search Appliance

FEEDFETCHER-GOOGLE – RSS reader fetcher



[A] GOOGLEBOT:

USER-AGENT: GOOGLEBOT

The Google crawler which reads your web page.

a) File Formats read by GoogleBot:

- Adobe Portable Document Format (pdf)
- Adobe PostScript (ps)
- Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)
- Lotus WordPro (lwp)
- MacWrite (mw)
- Microsoft Excel (xls)
- Microsoft PowerPoint (ppt)
- Microsoft Word (doc)
- Microsoft Works (wks, wps, wdb)
- Microsoft Write (wri)
- Rich Text Format (rtf)
- Shockwave Flash (swf)
- Text (ans, txt)

Google is always striving to get its hand on as much relevant content as it can. One of the file formats that is getting a lot of eye balls on the web in the last couple of years is the Flash files (.swf). *Of late, there is a lot of evidence on the user blogs & SEO forums that indicate that Flash files are also being read & indexed.*

- b) **File Formats avoided by GoogleBot:** Some file extensions have a very large file size and are considered untouchables by the Bot. Some of these are - exe, dll, zip, dmg etc.
- c) **Guiding GoogleBot:** GoogleBot can be directed to crawl a certain page through the use of Robot.txt file (which uses the agent name: user-agent) or through the use of a Meta Robot tag - This tag resides inside the website code & has the following format -

```
<meta name="googlebot" content="robots-terms">
```

robots-terms

- **noindex**
Document will not be indexed by Googlebot.
- **nofollow**
Internal and external links in the document will not be followed by Googlebot.
- **noarchive**
Google will not archive a copy of the document (Google's Cached Page).
- **nosnippet**
Google will not display snippets and will not archive a copy of the document (Google's



Cached Page). A snippet is a text excerpt from the returned result page that has all query terms bolded.

- If this Robots META Tag is missing, or if there is no content, or the robot terms are not specified, then the robot terms will be assumed to be "**index, follow**" (e.g. "**a11**") which is the default indexing behavior for most major search engine spiders.

GOOGLEBOT has got two assistants to take away some of its load:

- FRESHBOT:** It is a relatively newer Bot which is used to crawl the updated pages on the web. It crawls the pages which are already in the index and looks for their updated versions. Thus, it makes sense to update your site as frequently as possible.
- DEEPBOT:** This guy follows every link that it could find & download as many pages as it could. It's a crawl maniac to say the least. But, this brings a better and holistic picture of the page to Google. In this process, Google also gets a more complete picture of the composition of a site. This Bot usually arrives at the end of the Google's monthly ritual of Backlink & content inspection (called **Google dance**). At this point of time, they take as much content as possible for a deeper level of Indexing.

[B] MEDIABOT:

USER-AGENT: MEDIAPARTNERS-GOOGLE

This crawler from Google is the lynchpin for serving contextually relevant ads to the Publishing sites. Its purpose is to analyze the content of the pages so that AdSense program can serve meaningful ads on these sites. This crawler should not be restricted to crawl the websites which are using AdSense. The webmasters can use this:

User-agent: Mediapartners-Google*
Disallow:

Topic of debate: *Running AdSense on a site helps with that site's ranking – As 2 crawlers – GoogleBot & MediaBot crawl your site resulting in better & frequent indexing?*

[C] IMAGEBOT:

USER-AGENT: GOOGLEBOT-IMAGE

Some of the known findings about IMAGEBOT are:

- The Imagebot scavenges the web for images to place in their image index.
- The ranking of Images for a particular keywords depends on certain factors – Filename, Surrounding text, Alt text & Page Title.
- If your website is not focused on Image inventory & download, then it makes sense to block Imagebot from crawling your site – using your robot.txt file.
- Blocking ImageBot also saves some bandwidth



[D] ADSBOT:

USER-AGENT: ADSBOT-GOOGLE

Some of the facts that have surfaced about this new member of the Google crawler family are:

- AdsBot serves a very specific purpose as far as crawling is concerned.
- It is geared to provide wisdom to the Adwords program of Google by -
 - By analyzing the content of the Landing pages related to an ad
 - This content analysis helps in determining the Quality Score for a particular ad
 - This Quality score in association with the Bid Amount & CTR (Click Through rate) is used by Google to determine the ranking score of an Ad for a particular Keyword.
- Thus, it makes sense not to block AdsBot if you are an advertiser & is using the Adwords.

[E] GOOGLEBOT-MOBILE:

USER-AGENT: GOOGLEBOT-MOBILE

Some known facts about the Mobile Content crawler from Google are:

- Google does use a specific crawler to gather mobile content
- Google indexes **public** mobile web content
- If your content appears to be available only to a subset of all mobile users (for example, only to subscribers of a certain mobile service provider), it may not be indexed.
- Users can search the mobile web on their mobile devices using Google Mobile Web Search.

[a] Getting Your Mobile content Indexed

The steps for this are roughly the same as for Non-mobile content -

- Submit Mobile Sitemaps to the Google Mobile Index just in the same way as the Non-mobile site maps are submitted.
- You create and add Mobile Sitemaps to your Google Webmaster Tools account in a similar way to Sitemaps for non-mobile content.
- If your Mobile site has changed, then you can resubmit your map



[F] GSA-CRAWLER:

The gsa-crawler is the search appliance robot that performs the crawling on a web site. The crawler identifies itself with every page it downloads from any web server by specifying a user agent that can be stored in a web server log file by webmasters.

- The Google Search Appliance uses standard Google rules, such as searching for all words and treating upper and lowercase letters the same. It recognizes the minus sign (-) to exclude unwanted words, but does not allow the word NOT.
- Google Search Appliance default search results look like the public search engine.
- The search results header includes the search field, search terms, and number of matches, and a suggested alternate spelling, based on the site dictionary, if appropriate.
- Each search result item has the title, URL, with the size and date if available, and a "snippet" from the document shows the matched term in context whenever possible.
- The Google Search Appliance is an excellent search engine for HTTP-accessible content.

The identifier used by the crawler consists of:

- The user agent name, which, by default, is set to gsa-crawler.
- A unique identifier that is assigned for each search appliance.
- The problem email address you entered in **Administration > System Settings**.

If you keep the user agent name gsa-crawler, the accessed web servers might see an identifier such as

`gsa-crawler (Enterprise; GID01065; yourname@yourcompany.com)`

The email is a required part of the identification to allow webmasters to contact you if the search appliance affects them negatively by crawling their sites too rapidly.

There may be pages or sites in your organization that you do not want the search appliance to crawl, such as password-protected directories with information that you want to keep private. To prevent the gsa-crawler from accessing the information on these servers, you can either:

- Enter their URL patterns in [Do Not Crawl URLs with the Following Patterns](#)
- Create and put a robots.txt file in the root of the server. A robots.txt file consists of the user-agent name and one or more lines of instruction for the robot.

For example:

```
# /robots.txt file for gsa-crawler (This is a comment line.)  
User-agent: gsa-crawler (This names the user-agent that the file targets.)  
Disallow: /*.cgi (The gsa-crawler will not be allowed to crawl any CGI files.)  
Disallow: /*.pl (The gsa-crawler will not be allowed to crawl any Perl scripts.)
```



Allow: /\$ (The gsa-crawler is allowed to crawl everything else.)

Disallow: / (This prevents the gsa-crawler from crawling *anything* on the site.)

(source: Search appliance documentation of Google)

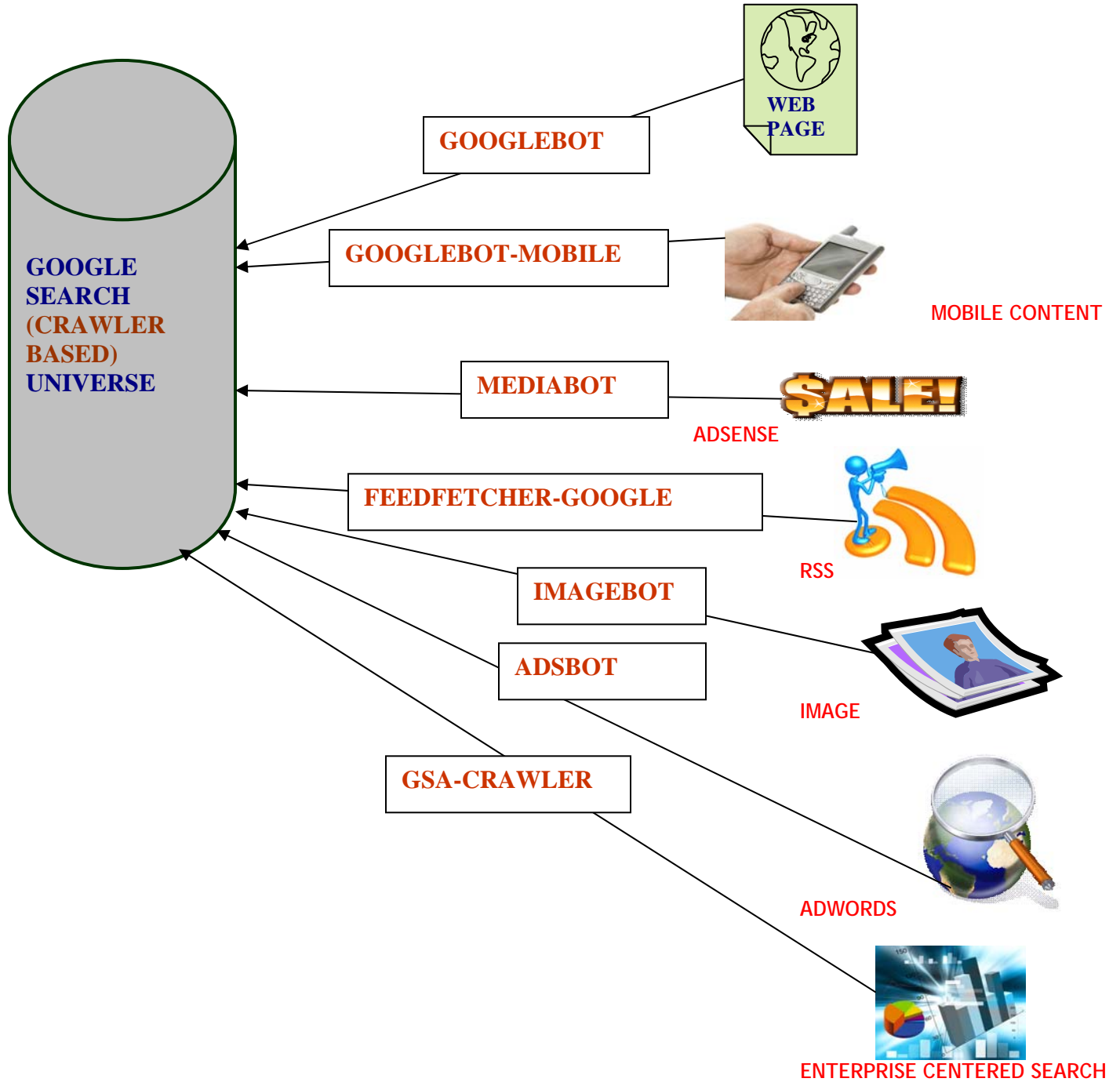
[G] FEEDFETCHER-GOOGLE:

This is the RSS & ATOM feed crawler of Google. The content that will be crawled by Feedfetcher are:

- All Blogs - published through Wordpress, TypePad, Blogger etc. - from all sources
- Blogs written in these languages apart from English -
French, Italian, German, Spanish, Korean, Brazilian Portuguese and other languages as well.
- Usually, the average crawl frequency for Feedfetcher is more than an hour - based on the frequency of your site update.
- So, if your Blog publishes a site feed in any format & pings an update service, then the contents of this feed will be indexed in the Blog Search.

So, If you are on Google Blog Search, rest assured that the Feedfetcher is doing all the collection task which makes the search easier for you.

Summary (Visual):





About the Author:

Vineet Singh is a teacher and practitioner in the Search Engine Marketing domain. He has been working in the IT industry for the last 15 years. He is a graduate from IIT Kharagpur & Executive Leadership scholar from Cornell University. He is also the COO of TeleZent (www.telezent.com), a research & Training firm in the Internet Advertising & Marketing domain.

Email: svineets@yahoo.com