

Trends in Blog Searching

[Christina K. Pikas](#)

Christina Pikas is a Technical Librarian at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. She works with engineers and scientists to resolve information needs in all areas of engineering, science, computer science, and mathematics. She received both her MLS and her BS in Physics from the University of Maryland. This is her first article in b/ITe.

Over the past few months there has been an explosion of information in conferences, on the web, and in library journals and newsletters about weblogs (blogs) and RSS feeds. b/ITe discussed RSS feeds and blogs in the [November/December issue](#) (pp. 3-12). In brief, blogs are reverse chronologically arranged collections of articles or stories that are generally updated more frequently than regular web pages. Just like any other information on the net, there is no guarantee of authority, accuracy, or lack of bias. In fact, personal blogs are frequently biased and can be good sources of opinion and information from the "man on the street." Because blogs can be updated on the fly, they frequently have unfiltered information faster from war zones and sites of natural disasters than the mainstream media outlets. Blogs are also good sources of unfiltered information on either faulty or very useful products.



In the beginning, blogs appeared in search results alongside regular web pages. Since blogs are not technologically any different from other web pages (that is, they are html, xml, javascript, etc. — it is their format, not their coding that is different), spiders and bots collect posts the same way they collect other online information. Search engines that place greater value on sites that are recently and frequently updated and are highly linked tend to rank blog posts very highly. Since the barrier to publication is so low in blogs, arguably much lower than for standard web pages, these high rankings were introducing a lot of noise into online searches. Odds are, you have run across several archived blog posts if you've searched on a controversial topic in the past year. Recently, most major search engines have altered their algorithms to push blogs down in the search results. Engines that only return two results from any one site use this feature to limit the impact of blogs on the search results.

Blog searching breaks down into at least two categories: information from within blogs/across blogs or addresses of feeds from blogs so that you may subscribe in your aggregator. Feeds and blogs are two different things, but are closely linked because most blogs have feeds and many feeds are generated by blogs. Just as in other web search tools, there are search engines and directories. At this time, blog search engines are where general search engines were before the Google Age: there are many competing smaller products, but no outstanding products dominating the scene. Searching for feeds was discussed in the [November/December issue](#) and will not be covered in depth here.

General Search Engines

Most of the larger search engines have changed their algorithms so that blogs are not the most highly ranked sites. My

recent testing of Google shows that they have made some changes to their engine so that the blog posts are generally pushed down to the third or fourth page of hits.

To use Google (<http://www.google.com/>) to find information across blogs, enter your keywords then `~blog inurl:archives`. Using `blog` and `archives` is somewhat redundant but using `archives` alone retrieves more newspaper articles and using `~blog` alone works and gets more hits but with less precision because "log" is apparently a synonym for `blog`. For example, a search on *hubble maintenance* yields forums in the second group of hits, and the first blog in the third group of hits. The search *hubble maintenance inurl:archives* yields six blogs out of the first ten hits. The search *hubble maintenance ~blog* yields seven blogs out of the first ten hits. The search *hubble maintenance ~blog inurl:archives* yields ten out of ten, but only 54 hits total. The `archives` term takes advantage of the fact that most blogging software automatically immediately archives all posts in files with the "archives" in the title and URL.

In Yahoo (<http://www.yahoo.com/>), adding `blog inurl:archives` to your search also seems to do the trick. In Teoma (<http://www.teoma.com/>), add `blog archives` to your search. The `inurl:` shortcut works only if used for the whole search phrase. A more simple search in Alltheweb (<http://www.alltheweb.com/>) yields similar results; type your keywords then `blog`. AltaVista (<http://www.altavista.com/>) is not recommended for searching for information across blogs. You might try adding `url:archives and blog*` to your search, but variations of this yielded no relevant hits for the Hubble telescope although hobbyist and war blogs were easy to locate.

Blog search engines

Unlike general search engines, many blog search engines do not crawl the entire web or even the full text of blogs. Instead, they crawl RSS feeds. There are several problems with this. First, not all blogs have feeds. Second, some feeds may be abbreviated or truncated and may not adequately represent the content of the full post. Some blogging software (like the popular Google-owned [Blogger](#)) only creates ATOM feeds, so these may or may not be picked up by specialty blog search engines. Finally, some of these search engines only crawl the feeds that at least one of their users reads so newer, smaller feeds may not be findable. Many of these search engines also track what's hot; that is, what subjects are occurring most frequently in recent posts. Ari Paparo posted a large list of blog search engines on his blog (<http://www.aripaparo.com/archive/000632.html>), so that list will not be duplicated here. Advertised blog search engines that only search registered blogs are also omitted. Instead, the several search engines below stand out for the way they combat the weaknesses listed above.

Bloglines (<http://www.bloglines.com/>) is a free (for now) online RSS aggregator. Bloglines does aggregate ATOM and RSS feeds so both are searchable. An advanced search is available with blocks equivalent to *and*, *or*, *not*. A search on *hubble maintenance* yielded a smaller number of results, but they were highly relevant. Bloglines appears to only search feeds to which at least one member subscribes. A directory of feeds is also available, but this is just an alphabetical list of feeds and is not very useful. Other aggregators also have search functions, so these may be the most convenient search tools.



Bloogz (<http://www.bloogz.com>) has some very nice advanced search features. For example, you may mark words that introduce noise with a ~ so they will be ranked lower but not excluded from the search. You may enter your keywords in whatever order and mark their importance to the search. Language searching is a helpful feature that could be key for looking for war blogs. The downside of Bloogz is that the current database seems rather small. Bloogz returned only eight results for the *hubble maintenance* search but they were very relevant and the results include a clip from the text.



Feedster (www.feedster.com) is unique in that it has blog-specific advanced search and most of its database fields are searchable using the list of fields available from the help page (<http://www.feedster.com/help/fields.php>). For example, if you want information from a blog that has in its description that it was written by a librarian, you might append to your search:

`weblog_description=librar*`. Additional uncommon advanced features like proximity searching, truncation, range searching, and synonym searching are also available. Feedsters offers RSS feed updates on your search.

Waypath (www.waypath.com) is one of the few full text search engines. Instead of just applying standard search technology to a subset of the web, Waypath uses content analysis to link related posts. It crawls the web and indexes the full text of blogs. You may enter the permalink for an interesting post and find related posts. Advanced search features are not provided, but the results are still very relevant. If searching for posts from other countries or in other blogs, search using terms from the local language or jargon. One of the most interesting features is that it offers a choice of two feeds for your search results. You can get the most relevant items, or the newest.

Summary

Blogs are everywhere and it is important to either be able to search them or make sure you're not searching them when you are looking for authoritative, accurate, and unbiased information. As blogs and RSS feeds either continue to explode across the net or start to go out of style, the search engines will adapt. Next time you are shopping for a new technology product try searching in a blog search engine to see what people are saying. Use Waypath to find related blogs. If you need to do a very precise search, use Feedster; but, if you want a little of everything, stay with the big general search engines (like the one that starts with G).

[Back to b/ITe Web page](#)