Searching More Effectively

ARTHUR WEISS, *AWARE*

Dear Arthur,

Whenever I search on the internet I get thousands and thousands of hits — most of which are irrelevant to what I'm trying to find. Can the internet be used for competitive intelligence without having to trawl through lots of useless information?

The short answer to this question is "yes." However, all I can do in a short column is give you some pointers on how to search more effectively.

First, consider why the information you are looking for would actually be placed in an internet-accessible site. Not all information will be online — in fact, much relevant information is still not available online. More important, even information that is online may not always be found through a search engine. Research estimates that over 80 percent of web content is not readily searchable using popular search engines such as Google or Yahoo! These "hidden" pages are often referred to as the "deep web" or "invisible web," and their content may be "hidden" for a number of reasons including:

- The relevant web page may be created dynamically or built from the content of a database.
- The web page may be new and therefore not yet indexed.
- The web page may be very large — search engines index only the first parts of a long web page. For example, Google is believed to index content of up to around 500K in size on some pages, although this is relatively new and an increase from the previous limit of 100K.
- There may be no text on the page or the page may be in a format that cannot be indexed. For example, Flash pages are indexed only by Exalead (a smaller European search engine).

WHY WOULD THE INFORMATION BE AVAILABLE?

There is a general question you should ask when looking for information: Why should the information be available at all? Answering this question can lead you to ideas on where the information will be held. Is there a reason why it would be online, or in a form that could be found by a search engine? If you believe that the information is likely to be found through a search engine, then you should use the answer to guide your search and your chosen keywords.

For example, the primary aim of a competitor's website is to sell company products and inform various stakeholders about the company. An academic site may publicize recent research. It should be relatively easy to find such material. In contrast, websites such as the Securities and Exchange Commission site (www.sec.gov) hold statutory corporate data and so have less reason to ensure that their material is easily searchable using a search engine — even though the data is available on the Commission's website.

UNDERSTAND HOW SEARCH ENGINES WORK

The next step is to think about how the search engines work, assuming that this is the route you decide on for finding the required information. Search engines need to show sites that match the search terms entered. These search terms are often (but not always) in the pages listed. Generally, the same terms will be in thousands of documents, so the search engine's next task is to rank the found documents by perceived relevancy to the search terms you entered. Search engines gauge relevancy by a mix of the following factors:

- The number of times the terms are in the document.
- The location of the terms in the document (for example, in a title or heading versus toward the end of the text, as well as in relation to each other).
- Links or referrals from other sites to the document — and their type and number. (These are effectively votes in favor of the page's quality.)
- Proprietary methods that are unique to and serve to differentiate each search engine.

From the search engine's perspective, it is crucial that the results shown link to pages that are the most relevant for the majority of users. From a website owner's perspective, the objective is to be found by a search engine. As a result, there is a tension between the site's objective (to be found, irrespective of site content or quality) and the search engine's (to show only quality sites that are relevant

to a particular search). This tension has led to the field of search engine optimization, which aims to help sites match the relevancy calculations of the search engines for particular searches.

However, this is not necessarily what you as a competitive intelligence analyst want. You want to find information that may not be optimized by the site owner to be in the top search results. Such information may not be aimed at the casual searcher, but rather at your competitor's suppliers, customers, or others who know where the information is held on the competitor's site. Instead, you get thousands of irrelevant (to you) results.

USE ADVANCED SEARCH ENGINE FEATURES

To overcome this problem, you need to think a bit like the site designers. How would what you are looking for be phrased on a web page? Would any words appear in the title of the page, or in the page's URL? Perhaps the page will not be an HTML page but an Acrobat document.

Instead of just entering key words, use the search engine's advanced search features to restrict or filter the search. As an example, if you think that the document is likely to be an Acrobat file with a specific word in the page title, you could search for "*intitle:keyword filetype:pdf*" on Google. This approach will give you documents with the key word in the title (indicating that this is what the page is about rather than just mentioning your key word in passing), eliminating 90 percent or more of the potential hits. You will also have a list of documents that are in PDF format — generally more content rich than the standard HTML page.

There are several such filters, and they vary across search engines. An easy way to learn about them is to read the various search engine help pages or use the advanced search options accessible from the search engine home page. Using them should make your searching much more precise, with fewer but more relevant hits.

DON'T RESTRICT YOURSELF TO GOOGLE

As a competitive intelligence analyst, you do not trust just one information source — you are prepared to check out other sources. This principle does not apply only to primary research interviews or even secondary research (looking at news articles, for example). It also applies to web searching.

Consider trying different search engines. There is remarkably little overlap between them. The results on Google may turn out to be completely different from those on Yahoo! or Ask. Each search engine uses different algorithms to calculate the relevancy of a site's content. A site that Google considers important for a particular search may be seen as unimportant by another search engine. The meta-search engine, Dogpile, conducted a study in July 2005 that suggested that only just over 1 percent of the top search results from the leading four search engines were the same. (http://missingpieces.dogpile.com/). The search engines tested were Google, Yahoo!, Ask (then known as AskJeeves), and Live.com (then known as MSN Search).

When searching, consider using all four. Two other search engines worth checking out are:

- Exalead (www.exalead.com), which has a number of unique search features unavailable on the others, such as wild-card searching.
- Northern Light (www.nlresearch.com), a recently reborn search engine, which gives you the option of searching news, journal articles, and other paid sources as well as business-related sites on the web.

SIDEBAR: CITED RESOURCES

www.aiip.org
www.archive.org
www.ask.com
www.completeplanet.com
www.dogpile.com
www.espacenet.com
www.exalead.com
www.google.com
http://infomine.ucr.edu
www.live.com
www.nlresearch.com
www.sec.gov
www.technorati.com
www.yahoo.com
www.resourceshelf.com
www.searchengineland.com
www.searchenginewatch.com
www.pandia.com

LEARN SPECIALIST AND DEEP WEB SEARCH TOOLS

Finally, to really benefit from the competitive information that can found through the internet, learn to use specialist and deep web search tools. Numerous tools can aid the competitive intelligence researcher — you can locate some of these by doing a search for "invisible web sources." Others can be learned from monitoring web sites on searching, such as resourceshelf.com, searchengineland.com, searchenginewatch.com, and pandia.com.

Blogs can be searched using the blog search options on Ask (www.ask.com), as well as sites such as Technorati (www.technorati.com). Historical web pages can be viewed from the "Wayback Machine" at www.archive.org. Patent information worldwide can be found from the European Patent Office at Espacenet (www.espacenet.com). There are also many sources for searching

databases and other material that is held in the "deep web." These sources include "Infomine" (a repository of scholarly and academic material at http://infomine.ucr.edu/) and a collection of over 70,000 searchable databases at Complete Planet (www.completeplanet.com).

EXPERT SEARCHERS

These tips should help you become a more efficient searcher, reducing the number of hits you get to a manageable quantity. However, there will be times when even these tips are not enough. If you think that the information will be available online but can't find it, then consider using an expert searcher to improve your search results. Such people spend their lives looking for — and finding — difficult-to-find information. The time saved will be well worth the money spent.

Speak to your company's information center, or consider outsourcing your search to a member of the Association of Independent Information Professionals (www.aiip.org), who should be able to advise you on how to find the information or find it for you.

[***Editor's Note:*** The internet and the web are not synonymous, although most users view them (incorrectly) as the same. The **internet** is a collection of interconnected computer networks that communicate using internet protocol (IP) and are linked at various levels by copper wire, fiber-optic cables, microwave connections, and so on. These networks carry information and services such as e-mail, online chat, file transfer, and the various documents and pages of the World Wide Web. The **web** is a collection of interlinked documents and other *resources* accessible via the internet.]

Arthur Weiss is managing partner with AWARE, a leading UK competitive intelligence consultancy. The answers given do not necessarily represent the views and opinions of either CI Magazine or SCIP. Arthur's blog at www.find-it-out.co.uk contains tips and ideas for better marketing and competitive intelligence practice. For more information, visit the AWARE website at www.marketing-intelligence.co.uk