

6 Navigation Tips for Terabytes of Data

Need to sift through terabytes of data? This article offers a step-by-step guide to using a search engine to find what you need.

This guide uses terminology, etc. from the search engine dtSearch®. dtSearch's enterprise and developer products can run "on premises" or on cloud platforms to instantly search terabytes across a wide range of online and offline data. Many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data can download a fully-functional 30-day evaluation copy from dtSearch.com

Step 1: Use the search engine to index the data. A search engine like dtSearch can search terabytes of "Office" files, PDFs, emails along with nested attachments, etc., even if you don't build a search index. But while unindexed search is slow, indexed search is typically instantaneous, even for multiple concurrent users across terabytes of data.

How do you get the search engine to build an index? Just point to the directories, email archives and other data repositories you want to index, and dtSearch will do the rest. In fact, the same index can include multiple different data repositories. For each data repository (including compressed archives inside of a data repository), dtSearch goes through every file, email and the like and automatically figures out the relevant file type.

In doing so, dtSearch uses information inside each file rather than relying on the file extension like .PDF, .DOCX, .PST, etc. It is all too easy to have an Access database with an Excel spreadsheet extension, or a PowerPoint with an email file extension. Looking inside the file itself to determine the correct format is essential to correctly parsing the data.

Step 2: Check the index log. Checking the index log is an important step to make sure that everything is fully indexed. For example, you can have "image only" PDFs mixed in with ordinary PDFs without even realizing it. An "image only" PDF is a PDF that may look ordinary, but is really just a picture only. When you try to copy and paste what looks like words, the copy and paste doesn't work because the underlying words are a pure image.

The indexing log flags image-only PDFs so you can run them through an OCR application like Adobe Acrobat to turn these into regular text-based PDFs. As a side note, when dtSearch updates an index, it need only look at what has been added, deleted or changed rather than rebuilding the whole index from scratch.

Article contributed
by dtSearch®

Step 1: Use the search engine to index the data

Step 2: Check the index log

Step 3: Leverage all of the many search options to refine your search query

Step 4: Sort and re-sort search results by relevance and other sorting metrics

Step 5: Generate a search report

Step 6: Consider caching

Step 3: Leverage all of the many search options to refine your search query. The main tip here is not to limit yourself to natural language unstructured search requests or simple word and phrase searching. dtSearch has over 25 different search types, everything from Boolean, proximity and concept searching to metadata-focused options and credit card recognition. Use the full range of search features to generate a query tailored to exactly what you are looking for.

One specific search option to keep in mind is fuzzy searching, adjustable from 0 to 10 to sift through minor typographical or OCR errors. If you are looking for coffee and it is misspelled coffre in an email or as a result of a blurry OCR'ed original, a low level of fuzzy searching will still pick that up. Fuzzy searching works on top of other search options.

The search options mentioned here, including fuzzy searching, work not only with English text but any of the hundreds of international Unicode-based languages. For use in a multi-user concurrent-searching environment, each search request runs on its own thread. That way, each user's search requests can proceed separately and with instant response.

Step 4: Sort and re-sort search results by relevance and other sorting metrics. After a search, dtSearch shows a full view of each retrieved file, email and the like with highlighted hits. If your search retrieves only a small number of files, scanning all of them is relatively straightforward. But when a search retrieves a large number of items, sorting becomes important.

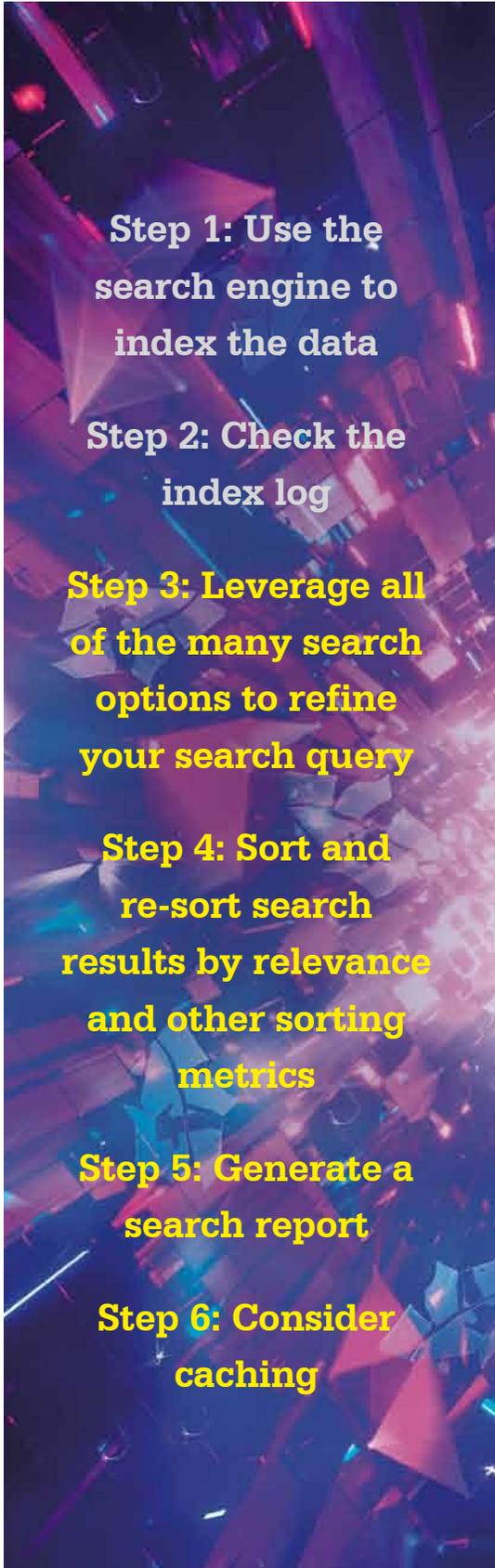
Relevancy-ranking uses a "vector-space" algorithm to sort by hit term density and rarity. If a search term is less frequent in indexed data, it will get a higher relevancy ranking. Say you search for coffee or tea. If there are millions of tea references but only a few coffee mentions, coffee references and especially files with denser coffee mentions will have a higher relevancy rank.

But the main point is that you are not stuck with your initial sorting. If you have relevancy-ranking as the default, you can instantly re-sort by descending or ascending file and email date, by file or email location, etc. Different sorting options can give you a better window into search results.

Step 5: Generate a search report. dtSearch can also generate a search report pulling together each hit across all retrieved files with as much context around each hit as you want. A search report is a great way to bring together a lot of hits across a large number of files into one easy-to-read summary.

Step 6: Consider caching. For caching, you need to go back to step one, where you build an index. Caching can store a full copy of each indexed file or email in the index itself. That way, when dtSearch goes to display a file, even if the original is no longer there or is subject to a spotty online connection, the search results can nonetheless instantly show the retrieved item with highlighted hits.

If you have terabytes of data you need to navigate, you are welcome to download a 30-day evaluation version from dtSearch.com to get started now.



Step 1: Use the search engine to index the data

Step 2: Check the index log

Step 3: Leverage all of the many search options to refine your search query

Step 4: Sort and re-sort search results by relevance and other sorting metrics

Step 5: Generate a search report

Step 6: Consider caching