

Fine-Tune Enterprise Search Parameters

The US military adjusted its search parameters for objects above US airspace following the Chinese balloon that traversed the country. That got me thinking about the need to fine-tune enterprise search parameters generally to make sure that items don't "fall through the cracks." But first, some background on how enterprise search works to put the search parameters in context.

Enterprise search such as dtSearch® instantly searches terabytes after first building a search index. Getting the enterprise search engine to index data is easy. All you have to do is point to the folders, email archives and the like to cover and the search engine will do the rest. In building its index, the search engine approaches each item in its binary format, as opposed to pulling up each in its associated application.

Before identifying the text and the metadata in a binary format, the search engine needs to figure out the correct file format so that it can apply the right parsing specification. Parsing specifications can be hundreds of pages long, so it is essential to get this determination right. The file extension is not a reliable indicator of file type as it is all too possible to save a Word file with a .PDF file extension or a PDF with a .DOCX file extension. Therefore, the search engine has to look inside the binary format for this critical information.

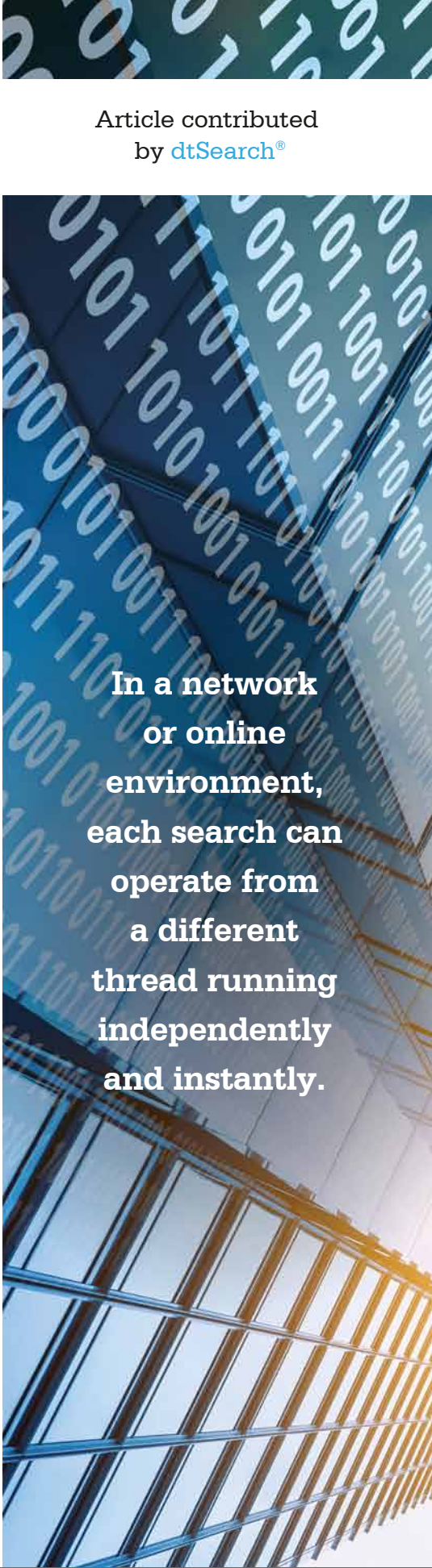
So what types of data can the search engine parse? All popular file formats, like PDF, Microsoft Word, Access, Excel, PowerPoint and OneNote, email formats like Outlook or Exchange, and even web-based data formats. The search engine can also automatically sift through compressed formats like RAR or ZIP and multilevel nested formats. That way, if the search engine comes across an email with a compressed attachment that includes a PowerPoint with an Excel file embedded inside, the search engine can still correctly handle the whole thing.

After recognizing each file format and identifying all the text and metadata, the search engine can go ahead with its index. After indexing, instant search across terabytes can proceed on an individual or a multiuser concurrent search basis. In a network or online environment, each search can operate from a different thread running independently and instantly. Instant concurrent search can even continue when an index updates itself to reflect new content.

But while indexed search is an unlimited commodity, the search parameters are key. Enterprise search has over 25 different search features that individual or concurrent users can choose from. Following are some tips for optimizing search parameters.

Tip #1: There are two basic types of search requests. The first is an unstructured natural language search request like an "any words" or "all words" search request. Here you would just enter some words: *ProjectABC contractor pipeline Dallas*. Generally speaking, however, a structured word and phrase Boolean and/or/not search request in combination with a proximity search request can better hone in on the

Article contributed
by [dtSearch®](#)



**In a network
or online
environment,
each search can
operate from
a different
thread running
independently
and instantly.**

right data more efficiently, like: (*ProjectABC* and not *ProjectXYZ*) and (*contractor w/55 pipeline*) and (*pipeline not w/7 freight railroads*) and (*Dallas* and not *Houston*).

Tip #2: The next tip would be to take advantage of any metadata in files, such as limiting the above query to only files that contain *pipeline* in certain metadata, like a subject field. Concept search can expand a search request to synonyms that you may not have thought of like *developer* or *builder* or even *subcontractor* for *contractor*. Fuzzy searching adjustable from 1 to 10 sifts through typographical errors. That way, if *contractor* is mistyped as *contrantor* in an email, or shows up with a similar slight misspelling in a PDF from an OCR mistake, the search would still pick that up.

Tip #3: In addition to word-oriented searches, a search engine can also look for numbers or numeric ranges. For example, if the project has a project ID that might range from 44876 to 44932, that could be an additional element in the search request. Or the query can also include dates or date ranges like *February 15, 2019* to *March 11, 2021*. Date searches can also automatically pick up common date variants like *11/8/20*. The search engine can also look for specific credit card numbers that might be relevant to the project. Or in the absence of specific credit card numbers to search for, the search engine can simply automatically flag any credit card numbers that may appear anywhere in the data.

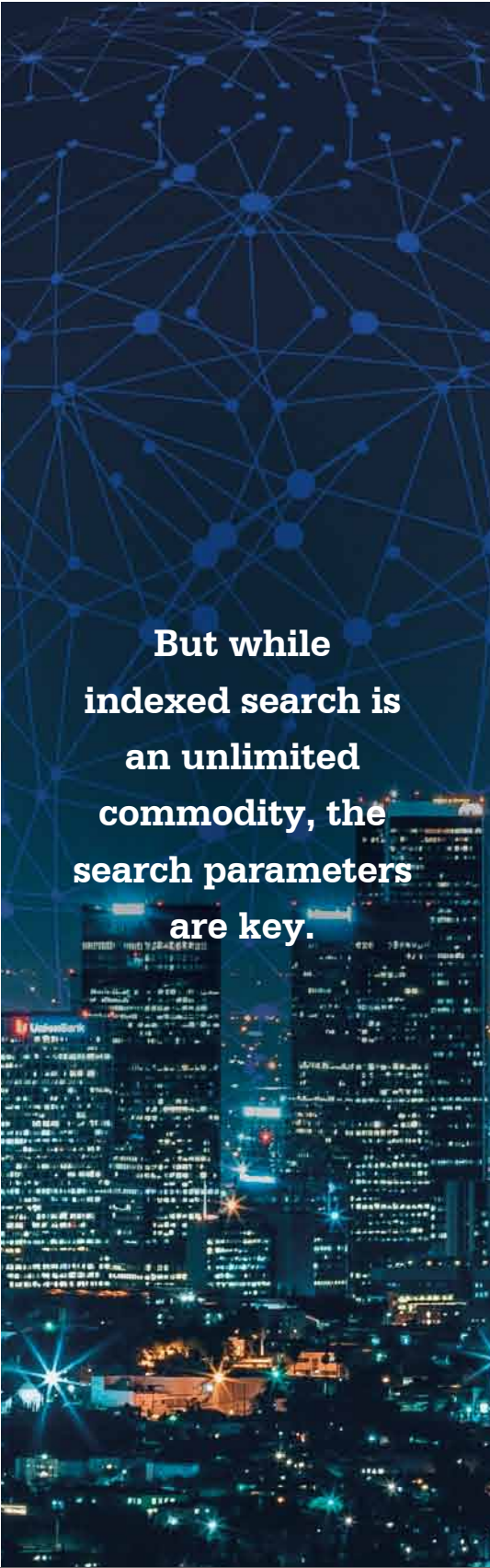
Tip #4: The next tip would relate to how a search engine prioritizes search result. By default, the search engine uses a vector-space relevancy-ranking algorithm to prioritize search results. What that means is if *contractor* is in millions of files but *ProjectABC* is just in a small number of files, then *ProjectABC* references would get priority, and files with the densest references to *ProjectABC* would get an even higher priority.

Beyond the default ranking, the search engine also supports variable term weighting. For example, the structured search request sample included *Dallas* and not *Houston*. An alternative would be to reconfigure the search request to not completely exclude files that mention *Houston*, just give them less priority, maybe giving *Dallas* a positive weight of 9 and *Houston* a negative weight of 6. Or you could extra-prioritize files mentioning *Dallas* at the top or bottom of a file, and extra-deprioritize files mentioning *Houston* in certain metadata.

Tip #5 and beyond: After looking at search results, consider instantly resorting by some other criterion like file date or file location for a different search results view. Regardless of the sorting mechanism, the search engine can display the full text of all retrieved items with highlighted hits for convenient browsing. Please visit the Features Map at [dtSearch.com](https://dtsearch.com) for even more search options and search tips.

About dtSearch. dtSearch has enterprise and developer products that run “on premises” or on cloud platforms to instantly search terabytes of “Office” files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch can instantly search terabytes with over 25 different search features, many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data to search can download a fully-functional 30-day evaluation copy from [dtSearch.com](https://dtsearch.com)

Article contributed
by [dtSearch®](https://dtsearch.com)



But while
indexed search is
an unlimited
commodity, the
search parameters
are key.