National System Administrator’s Appreciation Day: A SysAdmin’s Guide to Easier Workload

Article contributed
by dtSearch®

A system administrator’s job is not an easy one: keeping the network up and running, making sure the website stays online, managing new installs and upgrades, tracking data security, and so much more. But there’s one small part of these responsibilities that a search engine can help with: emergency requests to find a particular piece of data.

You know the drill. Suddenly, a corporate takeover back burnered a few years ago becomes red-hot again. Or maybe there are murmurs of litigation surrounding the original deal’s terms. Or perhaps an executive is simply curious about an archived email exchange on the subject. Whatever the reason, the system administrator must drop everything else and locate some critical piece of data, STAT. This is where an enterprise search engine comes in (instead of a “search the whole Internet for the right website” search engine like Google).

It’d be great if a system administrator could wait to install the search engine until this urgent request comes in and immediately find the critical item, but that is not how a search engine works. A search engine first needs to index the data before instantly searching terabytes. Now I know you’re probably thinking: isn’t indexing hard? Do we really want to add that on top of everything else the system administrator has to worry about?

You know the drill. Suddenly, a corporate takeover back burnered a few years ago becomes red-hot again.

Happily, indexing is no effort for the system administrator. All the system administrator needs to point to the folders, email archives, etc., to index and let the search engine take it from there. The search engine will automatically go through each item in all selected folders and figure out the file format for each, whether a PDF; a Microsoft Word, Access, Excel, PowerPoint, or OneNote file; an email format; a web-based format like HTML or XML; or even a composite format like ZIP or RAR.

After determining the appropriate file format of each item, the search engine will automatically parse all full-text and metadata and use that to build an index storing each unique word and each unique number in the underlying data. The index will also record the location of each unique word or number in that data. With that, the search engine can make available over 25 different types of search options.

Reprinted with permission of Spiceworks, spiceworks.com

While the system administrator can leverage these search features to find the right data, she can also extend these search capabilities to others. In fact, any number of individuals can concurrently search the indexed data, whether in a classic network environment or from a secure Intranet or public Internet site. An online search can proceed in a "stateless" manner, meaning that any number of individuals can instantly search terabytes without affecting each other's search threads.

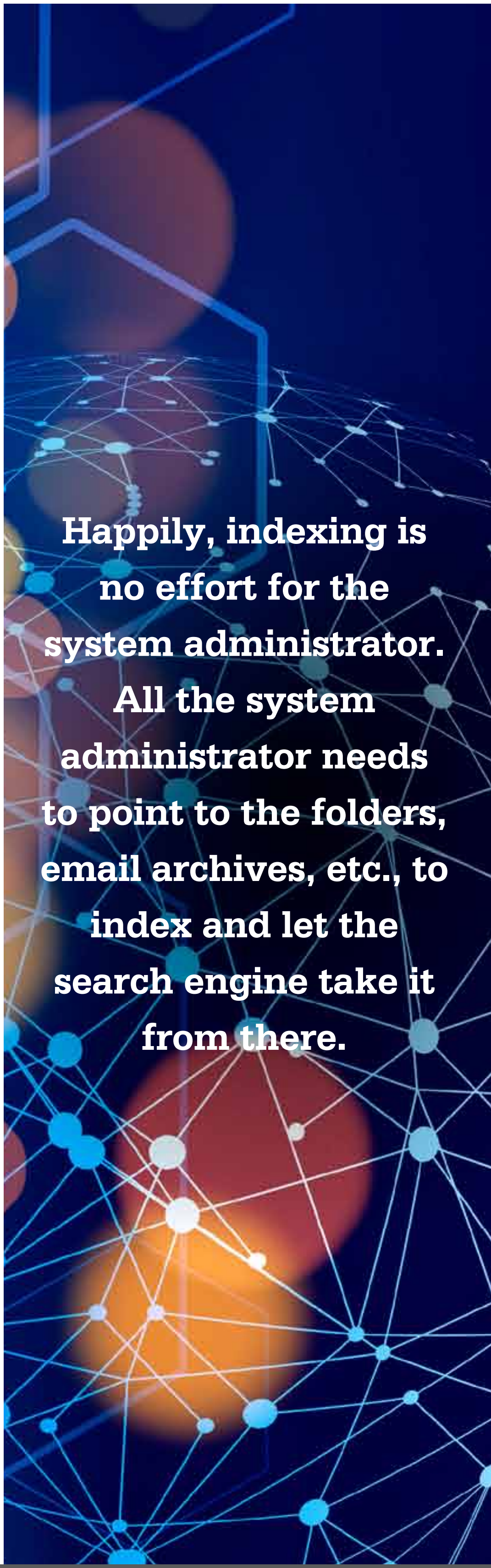
Let's look at a search request that might relate to our corporate takeover. At one end, such a search request might be a completely unstructured "natural language" series of terms: CompanyABC CompanyDEF takeover. On the other end, such a search request could be to a highly structured Boolean and/or/not and proximity query: (CompanyABC or CompanyDEF and not Company QRS) and (("board of directors") pre/46 "poison pill." ("Pre/46" here means within 46 words before.)

And the search query could add on a date or date range, like April 12, 2019, to June 6, 2020, finding anything in that range even if the date is in a different format like 1/19/20. Concept searching can find synonyms of any words using a general thesaurus or, of more relevance here, customized synonym rings. That way, if *CompanyDEF* was previously *DEFInc*, a search for *CompanyDEF* would also encompass DEFInc. Fuzzy searching can sift through typographical or OCR errors. If companyABC is mistyped CompaMyABC in an email, a low level of fuzzy searching would still find the "hit."

When a search turns out to be too broad, the system administrator can further refine it. For example, the system administrator could start with the above search request but limit the search to files or emails with Dallas or Chicago in key metadata. By default, the search engine will apply "vector space" relevancy-ranking to search results. With that, search terms less common in the indexed data will get a higher priority in search results. For example, if *Dallas* is common in the indexed data and *Chicago* mentions as relatively rarer, then *Chicago* would get a higher relevancy ranking.

For those that want to refine search results ranking further, a search engine can override the default relevancy ranking to give specific terms custom positive or negative weightings regardless of the default vector-space relevancy ranking. The search engine can even add special relevancy ranking to components of a search request that appear in specific metadata or close to the bottom or the top of the file. The searcher could also decide to immediately re-sort search results by a wholly different criterion, such as file date, file size or file location.

Article contributed by dtSearch®



Happily, indexing is no effort for the system administrator. All the system administrator needs to point to the folders, email archives, etc., to index and let the search engine take it from there.

Reprinted with permission of Spiceworks, spiceworks.com

After a search runs, the search engine can display a full copy of emails, files, and the like with highlighted hits for convenient review. The net result is that the system administrator looks like a genius for locating the missing data instantly. But as everyone knows, a system administrator cannot rest with just the basics. Some deeper-dive search engine notes follow.

"Non-conforming" data: This paragraph deals with all the data issues the system administrator does *not* have to worry about the search engine.

- This category includes files with "mismatched" extensions, such as a PDF file saved with a .DOCX extension. The search engine will figure out the relevant data type of each item by looking inside each binary file; a mismatched file extension will not affect that process.
- Also included in this category are multilevel nested formats. If an email has a ZIP or RAR attachment with an MS Word attachment and an Excel spreadsheet embedded inside the MS Word document, that is still no problem.
- Lastly, "hidden" text like black words against a black background or white words against a white background are all just text to a search engine.

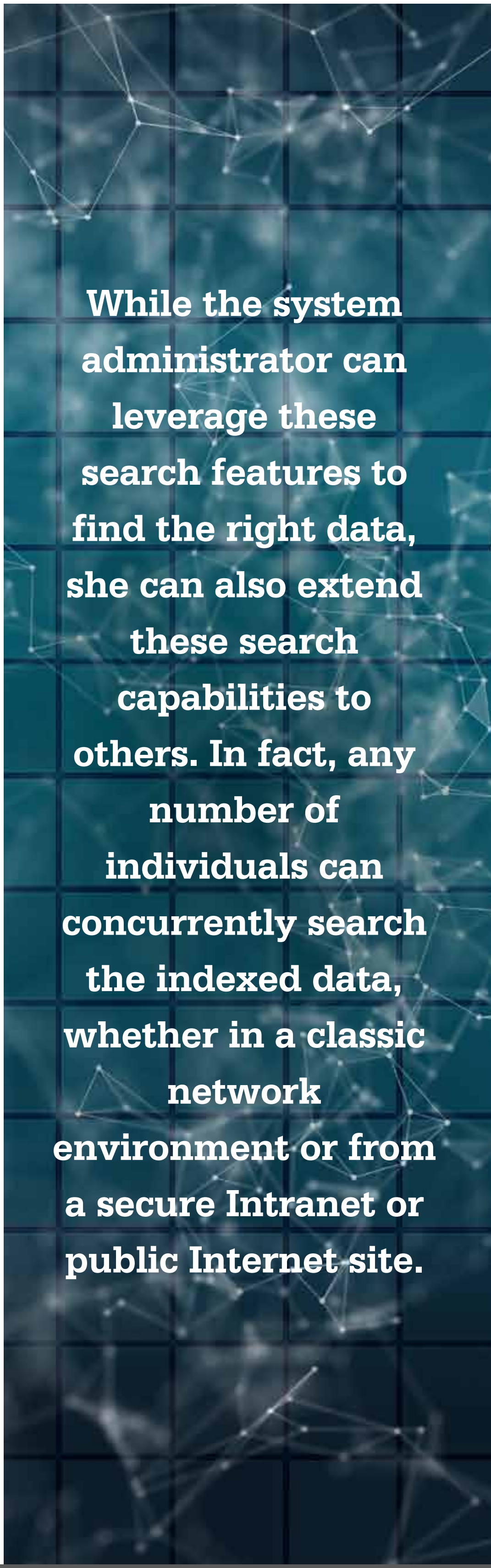
"Image only" PDFs: These are PDFs that look like ordinary PDFs from the outside but are really a picture only with no underlying text. The search engine can flag these at indexing time, letting the system administrator know that she needs to run them through an OCR program like Adobe Acrobat to make them full-text searchable.

Credit cards in data: The search engine can also flag certain data that shouldn't be sitting there on the network period. For example, the search engine can flag any credit card numbers in indexed data. With that, the system administrator can know where to find these unwanted additions to a data collection and take steps separately to remediate any such occurrences.

Updating an index: The system administrator can set indexes to automatically update as often as she wants without affecting continued concurrent searching. That way, the data can always be up to date, with no downtime.

One last thought for the system administrator. The search engine's document filters recognize and parse the different file and email formats. That same technology is also commonly built into third-party data leak prevention (DLP) software. Just one more thing to consider when the system administrator finds that critical takeover memo and moves on to her next task.

Article contributed
by dtSearch®



While the system administrator can leverage these search features to find the right data, she can also extend these search capabilities to others. In fact, any number of individuals can concurrently search the indexed data, whether in a classic network environment or from a secure Intranet or public Internet site.

Reprinted with permission of Spiceworks, spiceworks.com