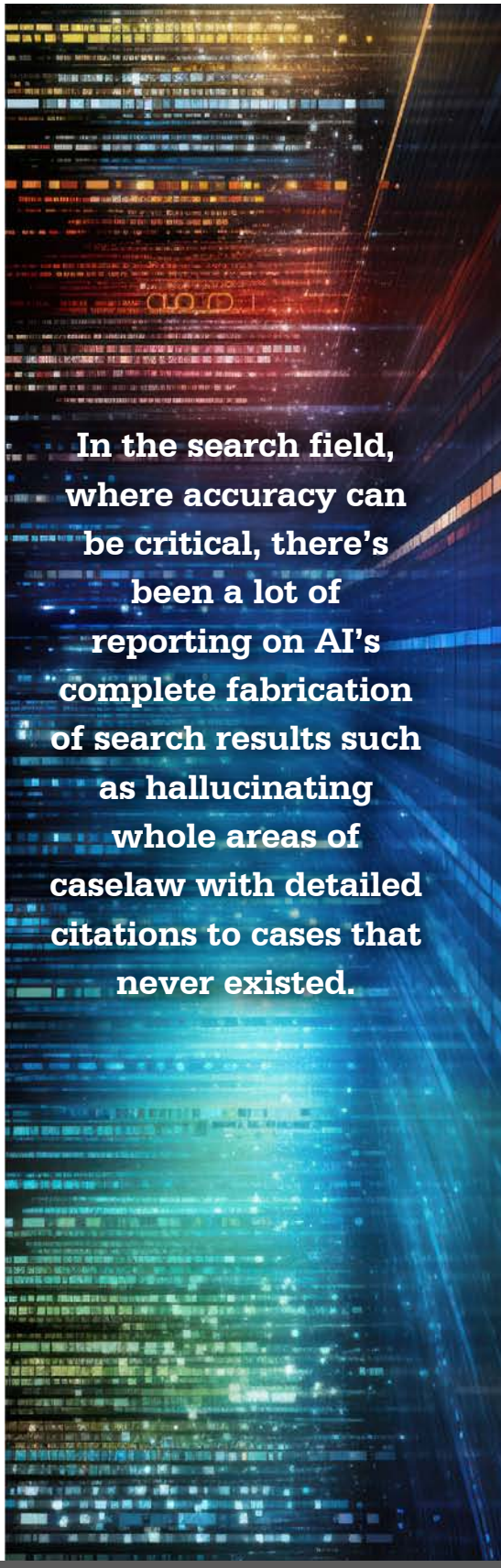
Precision Enterprise Searching in the Dawning AI Age

Article contributed by dtSearch®



For those who aren't familiar with dtSearch®, what does dtSearch do?

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

What is today's topic?

We get questions on how precision enterprise search generally and dtSearch specifically fit into the dawning AI age. Data mining and other dtSearch developer customers are increasingly looking at some type of AI tie-in for their products and they can do that with the dtSearch Engine for Windows, Linux or macOS. And even if an application doesn't require text search, the dtSearch Engine document filters can assist in the collection of AI training data.

Can you explain what the document filters do?

"Office" files, PDFs, emails, etc. typically look like gibberish in binary format. But from this seeming gibberish, the document filters can parse and extract the full text and metadata. The document filters can also parse and extract text and metadata from online formats along with Office365, SharePoint and other remote files that present as part of the Windows file system. And the dtSearch Engine document filters can work with databases like SQL or NoSQL covering BLOB files and associated metadata. All of this can be important AI training data.

Do you have to tell the document filters what types of files it is working with?

To apply the right parsing specification, the dtSearch document filters need to figure out exactly what file type each item is: PDF, Microsoft Word, Access, Excel, PowerPoint, OneNote, an email format, a specific web-based format, etc.

However, the document filters can do that on their own using information inside each binary file. A mismatched file extension won't matter here as identification doesn't depend on a file's extension.

How deep do the document filters go in parsing file data?

If an email has a ZIP or RAR attachment containing a PDF plus a Word document that itself embeds an Excel spreadsheet, the document filters will parse all text and metadata down to the innermost item. Viewing files in their native applications, you can click around extensively and still miss certain metadata. But all metadata is right there in the binary format for the document filters. Similarly, while you'd be hard-pressed to spot "hidden" white text against a white background or black text against a black background viewing a file in its native application, such text is readily apparent in binary format and hence accessible to the document filters.

And dtSearch can also work with multiple languages, right?

Yes. dtSearch supports Unicode which is the main text standard covering hundreds of international languages. These include not only European languages, but also right-to-left languages like Hebrew and Arabic, and double-byte-character Asian languages like Chinese, Japanese and Korean. A single file or email can have multiple successive Unicode-supported languages, and the document filters will follow the entire progression.

How do dtSearch's own enterprise search products fit in with AI?

AI is an amazing tool for problem solving and creativity, like generating an image of the present-day New York City skyline with larger-than-life dinosaurs on top of the skyscrapers, painted in the style of the French impressionists. But AI's creativity has both benefits and drawbacks. In the search field, where accuracy can be critical, there's been a lot of reporting on AI's complete fabrication of search results such as hallucinating whole areas of caselaw with detailed citations to cases that never existed. While we support dtSearch's developer customers looking at AI tie-ins to their own products, we are keeping AI separate from our own "off the shelf" enterprise search for now. Hallucination is the exact opposite of how enterprise search operates.

How does enterprise search work?

Enterprise search enables individual or concurrent instant searching across terabytes of an organization's data only after first indexing that data. Indexing uses the document filters to

Article contributed by dtSearch®

automatically go through the folders you select and record each unique word or number and the position of each in the data. An indexed search for *Large Company ABC* within 27 words of *Huge Company XYZ* with metadata including *New York or Texas and not Bolivia* will find only specific files, emails, email attachments, etc. that exactly meet those criteria.

What about less structured search requests?

dtSearch also supports unstructured natural language searching. Say you enter the names of all the planets in the solar system in an unstructured search request. If *Mercury, Neptune* and *Jupiter* have the fewest mentions in the data, then files containing any of these would get a higher relevancy-ranking, and files with the densest mentions of these would get the highest ranking.

And other search options?

Fuzzy searching adjusts from 0 to 10 to sift through OCR or typographical errors like an email with a mistyping of *plamet* for *planet*. dtSearch can also search for specific numbers and numeric ranges as well as dates and date ranges. A search for *date(1/5/24 to 2/15/24)* can pick up both *1/16/24* and *Jan 16, 2024* in the full-text or metadata. Other indexed search options range from hash value generation and search to identification of any credit card numbers across the data.

How do search results display?

dtSearch displays a complete copy of retrieved items with highlighted hits for convenient navigation. For a different window into search results, the software lets you instantly re-sort relevancy-ranked search results for example using a completely new criterion like file location or file date.

What about scalability across multiple end-users searching the same data?

While indexing is resource-intensive, searching is resource-light. Online search can even run in a stateless manner, accommodating any number of instant concurrent search threads. As data changes, dtSearch can automatically update indexes using the Windows Task Scheduler, with no disruption to continuing concurrent searching.

Final thoughts?

Looking for individual or concurrent instant precision searching across terabytes of enterprise data? Please go to dtSearch.com to download a fully-functional 30-day evaluation version.

Article contributed by dtSearch®