Like a Shot of Iced Espresso Through Enterprise Data: Metadata and Full-Text Search Together

Article contributed by dtSearch®

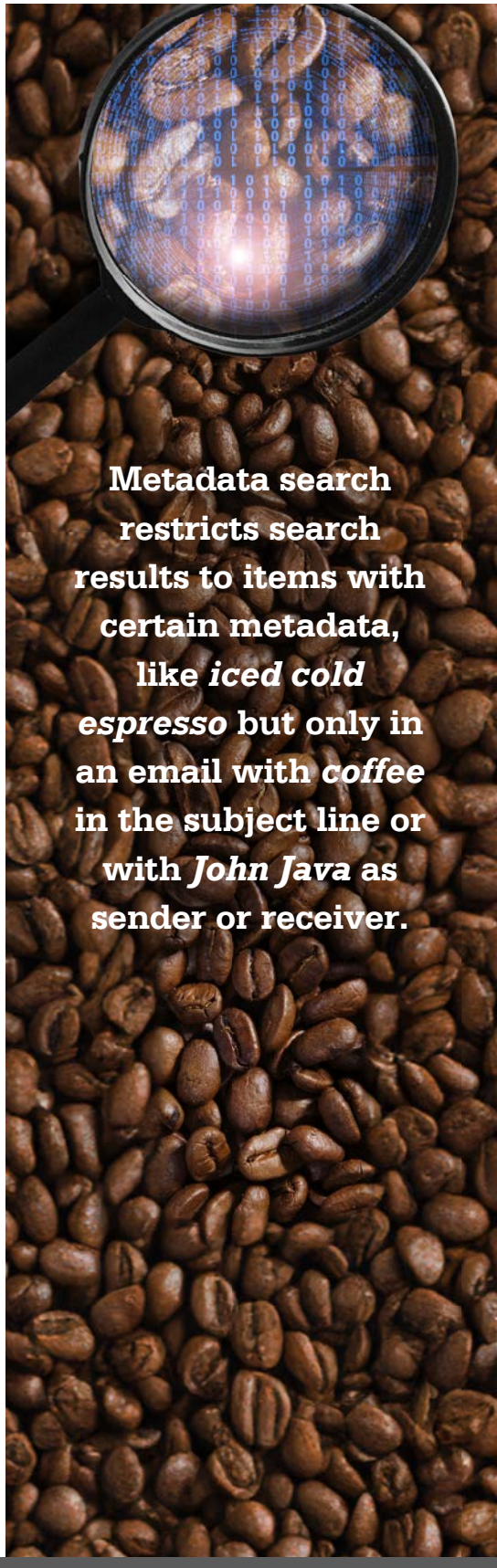
Combing through terabytes of enterprise data requires full-text search. Enterprise search like dtSearch® offers a wide array of full-text search options for instant multithreaded concurrent searching. But searches that target specific metadata are also important. And combining full-text and metadata search can be like a shot of iced espresso through enterprise data.

Instant searching through terabytes requires indexing as a preliminary step. Indexing may sound like a lot of effort, and it is for the enterprise search indexer. But all you need to do is check off the folders you want to index. The folders can hold local content such as files on a local area network, or remote data like SharePoint attachments and Office 365 files so long as these appear as Windows folder items.

To parse each file, the indexer needs to identify the file type: Microsoft Office or Office 365 Word, Access, Excel, PowerPoint, OneNote, Outlook/Exchange, PDF, XML, HTML, etc. The indexer can identify the file type based on the binary format; it doesn't matter if a particular file has a mismatched file extension like a Word document with a .PDF extension. The files themselves can be standalone or compressed in a ZIP or RAR archive. For example, dtSearch can automatically handle multilevel recursively nested files such as an email with a ZIP attachment holding an Excel spreadsheet with a Word document embedded in the Excel file.

Each index can hold up to a terabyte, and there are no limits on the number of indexes dtSearch can build and simultaneously search. The index architecture allows unlimited concurrent search threads, each of which can proceed independently to facilitate use in even very high-traffic situations. And the indexer can update indexes to accommodate file additions, modifications and deletions without interfering with ongoing concurrent searching.

A full-text indexed search covers anything in indexed data. Take a search for *iced cold espresso*. An "all words" query would look for items that contain all 3 words: *iced*, *cold* and *espresso*. An "any words" search would find items that contain even just one term. *An iced cold espresso* phrase search would retrieve only items that contain this exact expression. Proximity and Boolean (and/or/not) elements enable additional search refinement such as the phrase *iced cold espresso* within 12 words of *arabica beans* in a file that also mentions *cold-pressed* but not *chai* or *tea*.



Metadata search restricts search results to items with certain metadata, like *iced cold espresso* but only in an email with *coffee* in the subject line or with *John Java* as sender or receiver.

With concept searching, *coffee* and *espresso* would also retrieve *java*. Stemming finds different endings on the same root word like *press* and *pressing* in addition to *pressed*. Fuzziness adjusts from 1 to 10 to sift through typographical errors such as *espresso* mistyped or mis-OCR'ed as *esprexso*. Metadata search restricts search results to items with certain metadata, like *iced cold espresso* but only in an email with *coffee* in the subject line or with *John Java* as sender or receiver.

Metadata searching can also tie in with relevancy-ranking. Default relevancy-ranking uses a "vector-space" calculation. Take an "any words" search for *iced*, *cold* or *espresso*. If *iced* and *cold* appear in thousands of files but *espresso* pops in just a few dozen, then *espresso* files would get a higher relevancy rank. Items with the densest mentions of *espresso* would come out on top. But dtSearch also lets you customize relevancy ranking, like giving *coffee* in subject metadata an extra positive weight of 7, *John Java* as sender an extra positive weight of 3, and *John Java* as recipient a negative weight of 6.

Full-text and metadata date searching are both great tools for refining a search request. A full-text search for *August 7, 2023* would look for anything with that date. Or a search can look for *August 7, 2023* only in specific metadata like file modification date. Date searching can further automatically extend to different date formats and date ranges, with *date(July 31, 2023 to August 7, 2024)* picking up both *Aug 1 2024* and *8/1/24*.

The software can look for numbers or numeric ranges across everything or just in specific metadata. dtSearch can also identify certain number sequences like locating valid credit card numbers in data. Through Unicode, the software supports hundreds of international character sets, including European languages, right-to-left Hebrew and Arabic, and double-byte Chinese, Japanese and Korean. And the software can also find specific Unicode emojis, like the one that looks like a steaming coffee mug ☕ Still waiting on an iced coffee emoji.

Finally, the dtSearch Engine developer product can also search local or cloud-based SQL and NoSQL data, with integrated searching across SQL and NoSQL metadata plus associated files. Databases can reference external files or incorporate them as BLOB data. Such integrated searching allows for advanced data classification leveraging SQL or NoSQL metadata, full-text content, file metadata, or any combination of these. This integrated approach also enables faceted searching, letting end-users drill through layers of metadata to hone search parameters prior to running the search itself.

For a summer pick-me-up, sip iced coffee or tea. But for a data search lift, consider adding metadata elements to your full-text enterprise search.

Article contributed by dtSearch®