FEATURED PRODUCT

Document Filters & Enterprise Search

Most People Get The Idea Of Enterprise Search; Less Understood Are The Document Filters Underlying It

IF YOU LOOKED at a Microsoft Word file in binary format (as a search engine needs to review it), the file structure is so complex as to make it nearly impossible to pick out the text. In fact, MS Word documents include not only body text but also fields and often even hidden meta data. And MS Word files can have a nested structure, embedding multiple layers of other documents within the Word file.

Delving through these levels of complexity requires a programmatic implementation embedding a deep understanding of file structure. That is the job of document filters.

Document filters are a dynamic component. Every update, for example, that Microsoft makes to the MS Word format requires an adjustment to the document filters going forward, while still preserving backward compatibility with existing Word files.

One leading supplier of enterprise and developer text search software, dtSearch Corp., has spent over two decades building its own document filters. And the company continually upgrades its document filters to correspond with the release of new data formats.

In addition to Word, other MS Office file types that dtSearch supports include PowerPoint, Excel, Access, and OneNote. The document filters also support PDF, RTF, OpenOffice, HTML, XML, CSV, and many other file types, along with compression formats like RAR, ZIP, and GZIP/TAR. And the dtSearch document filters support recursively embedded versions of files, such as a Word file embedded in an Excel file contained in a ZIP attachment.

The dtSearch document filters can also support browser-compatible images in files, including recursively embedded files. The document filters further include Unicode support covering hundreds of international languages.

Document Filters: Not Just For Documents

With so much data now in emails, the dtSearch document filters also support email formats like MS Outlook, Exchange, and Thunderbird. And support extends beyond the email body and meta data to cover multi-layered nested attachments, including recursively-embedded images.

The dtSearch Engine APIs can also work with database data like SQL. While SQL itself is not a file format, it can include BLOB data consisting of embedded documents. The same integrated support for recursively embedded documents, meta data, images, and the like apply to this BLOB data.

Finally, the dtSearch Spider supports static and dynamic Web data (SharePoint, PHP, ASP.NET, CMS, etc.). Web data can consist of (or simply embed) document data such as HTML, PDF, XSL/XML, or even Office files, all of which require the document filters.

Beyond Document Filters: Hit-Highlighted Search

dtSearch enterprise and developer products can index more than a terabyte of data in a single index. A single index can span multiple file directories, emails and attachments, online data, and other databases. The products can create and search any number of indexes.

After indexing, the product line supports highly concurrent, multithreaded searching. Indexed search time is typically less than a second, even across terabytes of data. dtSearch products offer more than 25 search options.

For federated searching, dtSearch products support integrated relevancy ranking across both online and offline repositories. Following a search, the document filters enable hit-highlighting of federated search content.

In the dtSearch Engine, API filters and objects provide an even wider range of advanced data classification options. SDKs include native 64-bit and 32-bit APIs for C++, Java, and .NET (through current versions).

dtSearch®

Desktop with Spider

Network with Spider

Publish (portable media)

Web with Spider

Engine for Linux

includes 64-bit versions

Engine for Win & .NET

Document Filters
also available for separate licensing

Instantly Search Terabytes of Text

dtSearch

The dtSearch product line includes enterprise and developer text search products, meeting some of the largest-capacity text retrieval needs in the world. Please see www.dtsearch.com for hundreds of developer case studies and press reviews.

(800) IT-FINDS or (301) 263-0731 | www.dtsearch.com

PROCESSOR

Reprinted with permission from Processor magazine. For more information, visit www.processor.com or call (800) 247-4880.