The Difference Between Legal Search and Web Search

What You Should Know About Search Tools for E-Discovery

By Dr. Johannes C. Scholtes, President and CEO, ZyLAB North America LLC

Dr. Johannes C. Scholtes

Dr. Johannes C. Scholtes is president and CEO of ZyLAB North America and heads ZyLAB's global operations. Since Scholtes took over the leadership in 2002, ZyLAB has enjoyed double-digit expansion, consistent annual growth and profitability. Before joining ZyLAB in 1989, Scholtes was an officer in the intelligence department of the Royal Dutch Navy. Scholtes holds an M.Sc. degree in computer science from Delft University of Technology and a Ph.D. in computational linguistics from the University of Amsterdam. As of 2008, he holds the extraordinary chair in text mining from the Department of Knowledge Engineering at the University of Maastricht.

In many instances, when in-house legal professionals require advanced searching capabilities for e-discovery and legal activities, they often default to in-house variants of common Web search tools. However, Web search tools are not optimized for the types of activities associated with e-discovery, in large part because fundamental differences exist between the capabilities of Web search engines and the real search functionality and approaches needed to support the strategic requirements of legal, law enforcement and intelligence applications.

> ***"You need to know exactly how your search engine works and be able to explain it in court or to opposing counsel."***

One of the most compelling differences is that typical Web search engines are optimized to find only the most relevant documents; they are not optimized to find *all* relevant documents. Consider that with Web search engines, most companies and organizations place a premium on being found as close to the top of a search list as possible. Experienced users have become quite savvy in utilizing search engine optimization techniques to enhance high rankings. This level of sophistication works in both directions, though. People involved in criminal activities (such as fraud) don't want to be in the top 10 of a search engine result list, so they use advanced techniques to hide their documented activities and avoid appearing in any search list.

As a result, those searching in legal or law enforcement environments need to find all potentially relevant documents. Moreover, these investigators require different tool functionalities to quickly and efficiently navigate and review relevant document sets. The combination of these two requirements encompasses the practical difference between common Web search tools and legal search tools tailored for discovery-type activities.

An additional technical consideration is that, although Web search engines use many optimizations to continually perform real-time indexing of the Web, these optimizations come at a price: documents in non-standard formats will not be found, long documents will require a lot of time to review, and the processing of complex queries will be very slow (if even possible). Hit highlighting and hit navigation are often not available or operate too slowly. Moreover, with Web search engines, after documents are found, tagging them is not possible, nor can they be exported in a format required by regulators or courts.

Recognizing Different Search Capabilities

The strict e-discovery obligations and deadlines spelled out in the Federal Rules of Civil Procedures (FRCP) have highlighted the need for powerful in-house search technology, particularly in light of the current credit crisis. Meeting these requirements is becoming increasingly difficult given that the data repositories through which organizations now have to search for relevant and non-privileged documents are immense and ever-growing (i.e. they contain terabytes rather than "just" gigabytes of information). Given this context, consider the typical progression of e-discovery activities for most organizations: an organization receives a legal hold letter from a regulator or a third party; relevant custodians are established; and the organization's email and electronic files are handed off to a legal service bureau. The bureau would then find all documents that needed to be transferred to a third party in a specific, legally acceptable format. In some cases, these documents are in native file formats; in other cases, these formats are TIFF prints from electronic and paper files. The cost of such data processing services can be hundreds of thousands of dollars for a typical collection of, for instance, 250GB.

A remedy to alleviating these costs would be to implement an appropriate in-house search engine that could make a pre-selection of relevant documents and then create the document set that needs to be reviewed. However, many cases exist in which, rather than using an e-discovery-appropriate search tool, organizations implement Web search technology or Web search appliances to perform full-text searches on large email or electronic file collections throughout corporate networks. These organizations soon realize that the technical constraints of Web search technologies compromise the ability to meet set deadlines and address the requirements of regulators and courts, all of which can lead to higher costs and possible fines. Unfortunately, the limitations of Web search technologies are often not discovered until it's too late.

Understanding Search in E-Discovery

Searching is not only important for finding potentially relevant documents; it is also very important for supporting early case assessment activities. You must be able to quickly perform thorough and complex searches through your document repository, especially when you consider that searchers are under severe time constraints and/or are expensive investigators or (external) counsel.

ZyLAB has seen the most client "pain" when in-house legal teams and third parties confer to define the relevant search queries. As parties negotiate which documents need to be disclosed, lawyers establish what they consider the best Boolean, proximity and quorum operators needed to find specific data, and these operators are often combined and nested in hierarchical structures separated by brackets. Typical queries contain hundreds of words, and to catch spelling variations (e.g., from typos or optical character recognition [OCR] errors), a good search tool must be able to utilize wildcards (placeholders for beginning, middle and end of words) and fuzzy search (including support for first character changes).

Web search technologies are either unable to execute such queries or are too slow when attempting them. In these cases, executing a negotiated Boolean can take several days to finish, if it doesn't crash the system, so the query must be cut into smaller queries, with all spelling variations specified, which leads to an even more complicated search framework.

In addition, if a regulator or judge wants to verify that you have delivered all potentially relevant data, running additional fuzzy or wildcard searches might be required to find other documents. Cases are trending in this direction, and you need to make sure your in-house system can support it. You must be able to tag relevant documents or set them aside for deferred or external review, and you need to be able to show how you searched and what the results were.

Furthermore, your search engine needs to produce exactly the same results anytime it is used on the same data collection. Web search engines or engines based on certain high-dimensional statistical relevance ranking technology tend to produce different results over time. Cases relying on these kinds of searches are compromised in court.

Understanding Full-Text Indexing Processes

Most search engines use a "tokenizer" to enhance the searchability of data by removing punctuation and noise words, identifying words and determining character-set mappings (for foreign languages). This type of capability enhances your ability to perform the necessary full-text indexing of all relevant data. Of course, Web appliances can index for you, but their reporting and auditing functions may not match the standards required by regulators and the courts. With a Web search engine, you may not know exactly what data is in your index, and more specifically, what data is not in your index.

In some cases, a Web search appliance only keeps the 20,000 most relevant files in its index for a particular occurrence. This search engine is completely useless in a legal context. Furthermore, many Web search technologies cannot index documents that consist of compound documents (e.g., .zip and .pst), bitmap data, multimedia documents, older electronic file formats and encrypted files. If a legal search program runs into these types of documents, it will either separate them through a culling process or will automatically include additional processing to make such files fully searchable. When full-text indexing a document, document and file properties should be automatically extracted as well ("forensic indexing") and made searchable.

Remember too that crawlers need to automatically exclude corrupt, encrypted or unexpected file types, which can crash your crawler.

In sum, full-text indexing is a detailed process, and it illuminates the point that you need to know exactly how your search engine works and how to explain it in court or to opposing counsel. If there is existing case law that refers to the engine you use, that helps build your credibility as well.

Understanding Disclosure

Regulators, courts and opposing counsel often have very specific document format requirements for disclosed data. You should be able to support common legal file formats such as DII, EDRM XML, iPro or Concordance load files. Also, you should be able to redact documents, in which case you need to TIFF-print native electronic files to verify that all non-relevant information is no longer in the disclosed document. In addition, retrieved data needs to be collected and copied to a legal hold server, which is nearly impossible with Web search engines.

Failing to address the points mentioned above will lead to a lot of expensive and inefficient discovery work. Every irregularity, missed deadline or missing piece of data means a potential fine and more reliance on expensive outside vendors. Risk is diminished by understanding the required processes, matching procedures to those processes, using the right tools, and working with the right partners to lessen your exposure and costs.

For more information about standards and best-practices, consult:

- EDRM.net: provides the recognized standard for e-discovery;
- The Sedona Conference: offers reports on search, discovery, legal hold, records management and document production;
- TREC Legal project: evaluates high precision and recall search technologies; and
- www.zylab.com: showcases highly rated developer of award-winning e-discovery solutions used in high profile cases. ▮

Records Management, E-Discovery and Knowledge Management

ZyLAB's Universal Approach

Since 1983, ZyLAB has worked alongside professionals in the auditing, legal and intelligence communities to develop tools for investigating and managing large sets of archived data. These award-winning technologies have been bundled into the ZyIMAGE Information Access Platform, an integrated document, content and records management solution that enables businesses, auditors and legal professionals to capture, investigate, structure and disclose information in an efficient and secure manner.

ZyIMAGE helps you find more, giving you the proven technology required for comprehensive legal search:

- Support for large and nested complex Booleans, proximity and quorum search;
- Fast fuzzy (supporting first character changes) and advanced wildcard search (a*, *a, a*a, and *a*);
- Hit-highlighting and hit-navigation;
- Reproducible and reliable relevance ranking;
- Forensic indexing of file and document properties;
- Automatic language recognition;
- Indexing capabilities for compound objects such as nested emails, compressed files, email collections, databases and more;
- Extended index and search process auditing and reporting;
- Advanced visualization tools;
- Incremental indexing of live network data;
- Integration with records management, legal hold, identification, collection, legal review, (TIFF) productions and redaction processes;
- Advanced text analytics and machine translation; and
- A search engine mentioned in existing case law.