

## Complex Litigation & E-Discovery

### HARNESSING THE HIDDEN VALUE

#### Keyword-based vs. content-based searches in electronic discovery

By Jennifer Aira-Ventrella

When a company recently suspected its employees of embezzling \$50,000, it analyzed e-mail over a six-month period. A keyword search of the words bank account, transfer and similar terms noted little of any value. However, using a content-based search tool that identifies and subsequently organizes concepts into clusters, the company found an unusual number of messages between two employees discussing baseball using the terms dugouts and home runs.

Reviewers realized that the employees under scrutiny were not sports fans, but using baseball terms as euphemisms for their illicit activities. When investigators evaluated the dates of the bank transfers, they were able to compare them to e-mails describing a home run and other events, uncovering in 20 minutes documents that identified millions in fraudulent activity.

This success story is one of many where the use of traditional search methods involving keywords and Boolean connectors (e.g., “And” and “Or”) alone are giving way to emerging tools that are growing in popularity and helping to improve the dis-

*Aira-Ventrella leads engagement teams for BDO Consulting, a division of BDO Seidman, LLP in their Computer Forensics and E-Discovery Practice, specializing in large scale, multi-national electronic discovery projects.*

covery and investigative processes. According to the “Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery” section in The Sedona Conference Journal, “[t]here is a growing consensus that the application of linguistic and mathematic-based content analysis, embodied in new forms of search and retrieval technologies, tools, techniques and process in support of the review function can effectively reduce litigation cost, time, and error rates.”

As courts have become more familiar with electronic discovery, judges are also increasingly comfortable with keyword searches and their role in defining the scope of a project. This approach can be effective and successful for many types of projects and budgets. However, as technology becomes more sophisticated and projects become more complex, it is becoming obvious that keyword searching is producing results that can be both over and under inclusive due to false hits, recognition of noise words, improper use of Boolean operators by reviewers, and arguably most important, the “richness of the human language,” which makes it almost impossible for attorneys and paralegals to anticipate all of the words used contemporaneously in a case. Additionally, it is common for the identification of keywords to be an evolving process based on intelligence gathered from interviews and depositions conducted throughout an investigation. The need for

iterative keyword searches can add significantly to the overall cost of an investigation.

The baseball code example highlights the inherent risks and assumptions associated with keyword searching. Using code words may be rare, but people commonly incorporate unique acronyms into their conversations, particularly those online.

To aid navigation through electronically stored information (“ESI”), content-based reviews are increasingly used to reveal patterns a keyword-based reviewer may otherwise miss. Judge Facciola recently referenced in an opinion, “...recent scholarship...argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results,” *Disability Rights Council of Greater Washington v. Washington Metropolitan Transit Authority*, 242 F.R.D. 139. For example, a keyword search for embezzlement would identify any e-mail with that specific term, but a content-based search would highlight documents that have a substantive association with the messages uncovered by the search terms, including those that do not share any similar wording, (e.g., theft, fraud and money).

Content-based searching can be especially useful when allegations arise prior to litigation, particularly during internal investigations, when the in-house legal team will often have unanswered questions. “Search is an incredibly important component for e-discovery, but investigators are increasingly realizing that concept searching is a much more powerful tool for cases where individuals are trying to hide their activity,” said Ashley Watson, General

Counsel of Attenex Corp. Reviewers may have a vague mandate and not know the items for which they are looking. Content-based searches allow them to develop their focus and dynamically define the parameters of the project as they study the material.

Organizations with older documents that originate in hard copy format can also benefit from content-based searching. Taking advantage of existing technology and converting hard copy documents into electronic files using the optical character recognition (OCR) process, is clearly a better solution for large cases instead of manually reviewing millions of pages of documents. However, OCR technology is not perfect, thus identifying duplicates and subsequent keyword searching can be challenging. The ability to use a content-based search tool to group similar documents can potentially overcome these challenges and expedite review.

Since analysis is more sophisticated using content-based search tools and the goal is to increase the quality of the search results, keyword searching can be conducted in conjunction with the content-based search results.

Multiple methodologies exist to cull the universe of potentially relevant ESI. Simple techniques, such as filtering a data set by date, are frequently used. More recently, analytical techniques that identify and categorize messages by the sender and recipient are becoming more common in this e-mail-intensive era. Both techniques reduce review time by organizing files according to their relevance.

A more sophisticated mechanism that is gaining momentum is clustering, which is a statistical analysis of ESI that identifies relationships among documents that have similar concepts/content and clusters the identified relationships/documents together.

Other mechanisms involve further conceptual categorization. Taxonomy is a hierarchical approach to creating sets and subsets of concepts. Ontology, another technology similar to taxonomy, further expands the sets and subsets of concepts to identify relationships among those concepts. For example, in a litigation involving a corporation, a tool utilizing the taxonomy approach could potentially organize the data by personnel concepts (e.g.,

Professionals, Accounting Department, Human Resources), whereas the ontology approach would expand on the personnel types and identify relationships (e.g., under Professionals, Managing Directors would be identified to have a history of working with specific Directors and Managers).

Additionally, some tools have the ability to identify exception documents (e.g., password-protected, embedded objects), as well as identify and categorize foreign language documents.

Review platforms that implement content-based searching methodologies usually offer other advanced functionalities, such as de-duplication, identification of near de-duplicates (e-mail and/or user files), email analytics and date filters, which when used in conjunction with content-based searching, can further cull the volume of ESI, yielding the most relevant ESI for review. In the introductory baseball embezzlement case example, using date restrictions for the period of the suspected wrongdoing in conjunction with the clustering methodology, and later limiting the search to the two prime suspects, enhanced efficiency and accelerated the process of understanding the embezzlement scheme and quantifying its impact.

These options can help simplify the discovery process for outside counsel, who continue to lead discovery and are their clients' representatives in judicial proceedings. Accounting and consulting firms, as well as government and regulatory agencies, are recognizing the potential and seeing the value in these tools, as evidenced by their growing acceptance in the marketplace.

The still undecided question about whether courts will accept these content-based searches is the biggest hindrance to widespread adoption. Legal teams are primarily concerned about whether the results will withstand a challenge, although there are little to no references in the reported case law regarding alternative search methodologies.

The general lack of familiarity with the technology that supports content-based search tools also hinders its adoption. Aside from the mistrust of the automated nature of identifying concepts, there is confusion about the mechanics of the operation. Finding all documents that feature a particular word or phrase is a

notion readily understood. Identifying items that are similar in tone, idea or nuance is a grayer area that is more difficult to grasp.

Moreover, the more sophisticated technology is more expensive; sometimes causing the use of such technology to simply not be a cost-effective option. With that said, every electronic discovery case is different and content-based searching is not "one-size-fits-all." The benefits of the functionalities — de-duplication, date filters, concept searches and keywords — used synergistically can lead to a more effective, cost-efficient review, especially for larger matters. However, the tools employed in discovery on every case should be selected with an appreciation for the tools' functionality, balanced by the needs of the case at hand.

The universal goal of those involved with ESI is to bring order to modern discovery. The various methods for culling information are a logical evolution in the management of the exponential growth in ESI. The alternative content-based search methodologies are also a direct response to litigants striving to mitigate costs and develop strategic workflows to achieve the most relevant results in a cost-effective manner.

In a world where methodologies are ultimately judged on a reasonableness standard, it is more compelling to be innovative and find illicit activity of employees speaking in baseball code, than to strike out with keywords that may fail to identify the biggest issues of concern. ESI review technologies will continue to evolve and new methodologies — such as content-based searches — will continue to require experienced human oversight to take advantage of their search capabilities, manage costs and present results clearly and compellingly to management, counsel, enforcement personnel and the courts. ■



**BDO Consulting**  
A division of BDO Seidman, LLP