Leveraging Content Analytics to Reduce E-Discovery Risks and Costs

1. Introduction

"Electronic discovery"[1] - a concept that might have barely registered for many corporate counsels and law firms a decade ago – has today become a central issue in litigation, investigations and regulatory response activities. In fact, according to a recent survey, "electronic discovery [is] the number one new litigation-related issue for companies . . ."[2] This is not surprising when one considers the rapid growth of the e-discovery market (tripling in size from 2004 to 2007)[3] and the headline-grabbing cases (such as Morgan Stanley's $1.5 billion discovery-related debacle)[4] that are at least partially responsible for driving such growth.

But, beyond the headlines, what can organizations do today to manage the risk and costs associated with electronic discovery ("e-discovery")? What tools and practical techniques are available to help corporate counsels and law firms take control of the electronic discovery process? This Brief focuses on a key aspect of e-discovery - namely, the use of content analytics as part of the e-discovery process. More specifically, this Brief provides a high-level overview of content analytics; discusses the value it can bring to the e-discovery process; and provides key considerations for organizations evaluating or adopting content analytics for e-discovery.

2. Content Analytics Defined

Content analytics (in the context of e-discovery) refers to a set of computer technologies and processes designed to automatically determine and communicate the "meaning" of information to a human user. Content analytics is designed to help humans understand and evaluate information based on a computerized analysis of the information - versus what a human "tells" the computer about the information. The primary value of content analytics is that it can help organizations organize and analyze large volumes of unstructured information more efficiently than processes that are exclusively human-driven.

A very simple example of content analytics is found in some desktop programs that can automatically "recognize" dates, phone numbers, URLs, and other types of information as dates, phone numbers, and URLs, without any intervention on the user's part. The pattern matching algorithms and techniques used in content analytics are typically much more sophisticated and use a variety of techniques to classify and categorize information, such as the following:

WHERE LAW & TECHNOLOGY MEET

KAHN
CONSULTING INC.

June 2006

- **Context**. Identifying the relationship between documents, such as one document being a part of another, and therefore having contents or topics that are closely linked (e.g., an email message and its attachment). Also, identifying and grouping together documents that contain similar types of information (such as contracts, non-disclosure agreements, service level agreements, for example).

- **Concepts**. Identifying and grouping together documents that are conceptually similar based on the occurrence and frequency of nouns and noun phrases they share in common.

- **Visualization**. Presenting information regarding the content of a document - and its relationship to other documents - using visual techniques that allow document reviewers to quickly assimilate that information and make decisions regarding the documents.

- **Keywords**. Although not exclusively associated with content analytics, many content analytics tools also enable sophisticated search queries using keywords.

> *"[A] federal judge has ordered the department to do an accounting of the trust fund going back to the 19th century, when it originated . . . Interior had estimated that the larger search would cost $2.4 billion."*
>
> *House Panel Narrows Search into Payments Made to Indians, Wall Street Journal, July 15, 2002*

3. Leveraging Content Analytics

> *"Some respondents expressed more concern over the costs of litigation than they did over winning or losing lawsuits."*
>
> *Litigation Trends Survey[3]*

If designed, implemented, and used properly, content analytics can offer several improvements to the e-discovery process - a fact that the drafters of the Sedona Principles for Electronic Document Production[5] noted when they stated that "a principal advantage of electronic information is that high-speed methods exist to determine the existence of patterns of words, thereby allowing the narrowing of searches for relevant information."[6] The benefits of content analytics are explored further in this section.

Reducing review costs

The cost of human review of information for relevance and privilege often comprises the greatest single line item in the overall costs that organizations face during e-discovery. For example, in a recent case, the potential cost of just *the privilege review* was estimated at between "$16.5 million and $70 million.[8] Other than seeking a reduction in hourly rates charged for legal review of collected materials, there are two primary options for reducing this category of cost:

1) Reducing the number of documents that need to be reviewed, and
2) Increasing the efficiency of the review process.

Content analytics can help organizations achieve both of these goals, as explored in more detail below.

Reducing volumes

Electronic discovery is "the electronic equivalent of finding a piece of paper buried in a warehouse. . ."

The National Law Journal[9]

The volume of information that is potentially responsive to large cases can be staggering, and the problem of volume alone can play a key role in the strategy, cost, and outcome of such cases. In 2004, the volume of email messages sent by businesses worldwide exceeded 1 exabyte for the first time[10] (to put this in perspective, it would take over 12 million desktop computers to store an exabyte of data).[11] Beyond email, a standard desktop computer today can store the equivalent of 40,000,000 typewritten pages of information.[12] Large e-discovery exercises can involve the review and production of millions of pages of documents found in corporate messaging systems and other systems across the enterprise.

Content analytics can help to reduce costs by reducing the number of documents that must be reviewed and managed. Content analytics and related filtering, de-duplication, and suppression (i.e., not exposing irrelevant documents to reviewers) techniques can help to ensure that only the matter's responsive information is moved into the review and processing workflow, thereby reducing costs.

Reducing time required

Content analytics can also be used to reduce the amount of time spent processing, reviewing, and producing documents. As mentioned above, lowering the volume of documents requiring review will help in this regard. In addition, by facilitating the topical grouping of documents and presenting that information in a visual form, content analytics can enable reviewers to more quickly evaluate the documents they are presented with.

Moreover, by enabling organizations to more quickly understand the body of documents that is responsive, and the nature of those documents, content analytics can assist attorneys in assessing matters and developing legal strategies earlier in the process.

"[E]ven to this day, neither side to this motion has demonstrated to this Court a complete mastery of what types of documents were generated... how they were used, or their significance. The result was inevitable: discovery proceeded at a breakneck pace, and information was received faster than the attorneys could absorb it. As a result, both sides were the losers. They lavished huge sums of time and money on an issue that did not remotely justify the expenditure, and which would have been more profitably spent focusing on the merits of this case."

Danis v. USN Communications, Inc.[13]

WHERE LAW & TECHNOLOGY MEET

KAHN
CONSULTING INC.

4. Considerations for Evaluating Content Analytics Capabilities

Organizations evaluating or adopting content analytics tools as part of their e-discovery strategy should ensure that such tools are adequate to address their requirements. There are a multitude of issues that organizations conducting such assessment should consider, many of which are outside the scope of this Brief. However a selection of key considerations is discussed below.

> *"Regardless of the method chosen, consistency across the production can help ensure that responsive documents have been produced as appropriate."*
>
> *Sedona Principles*[14]

Consistency and Quality

Organizations should seek evidence that content analytics tools consistently perform to an expected level of quality. The complexity of content analytics technology and algorithms may make it difficult to directly evaluate the quality of the underlying technology, but organizations should consider pilot programs, testing, and other quantitative approaches to assess the consistency and quality of the tools used and their output.

Native file review and production

Organizations should ensure that the process used to review documents preserves the original information in its unaltered form and should be prepared to produce information in that original (or native) format. Although courts have generally indicated a willingness to accept non-native formats for electronic information (e.g., such as PDF and TIFF), there may also be occasions where producing the native format may be advisable or necessary - especially when dealing with complex and compound documents (such as spreadsheets with formulas and links to external data, for example).

In any case, as it relates to litigation, issues regarding the format of production should be addressed with the opposing side prior to commencing the production process. In fact, discussions on this topic are contemplated in the proposed new Federal Rules of Civil Procedure, which direct parties to discuss, "any issues relating to disclosure or discovery of electronically stored information, including the form or forms in which it should be produced," early in the litigation process.[15]

Additional considerations and capabilities

1) **Group categorization.** The ability for document reviewers to assign document classification information or other metadata to documents that have been grouped together through concept matching, keywords searches, and so on.
2) **Sophisticated filtering.** The ability to exclude information from the review process based on sophisticated filters that target metadata such as file type, creator or custodian name, date ranges, and so on.
3) **De-duplication.** A high percentage of information typically collected during the e-discovery process is often duplicates. According to one estimate, 70% of all corporate email messages and documents are duplicates.[16] In any case, eliminating duplicates from the review process will typically provide a significant reduction in the overall volume of the information
4) **Suppression**. Because there may be occasions where duplicate information may be required during the production phase of e-discovery, it may be preferable for content analytics tools to "suppress" duplicate (or filtered) information from the

reviewer, rather than deleting or otherwise removing such information from the overall review pool.

5) **Annotation capabilities.** The ability to add categorization information and other notations to reviewed documents without materially altering the original document.

5. Conclusion

E-discovery is critical to the way that organizations manage themselves and their digital information during normal business operations and in connection with litigation, investigations, and audits. The amount of time, money, and resources expended on e-discovery can be staggering for those organizations that are unprepared. Content analytics is a tool that organizations should evaluate and consider as a key weapon in helping them better survive and even win the e-discovery battle.

> *"Nearly 90% of US corporations are engaged in some type of litigation, and the average company balances a docket of 37 lawsuits. For $1 billion plus companies, the average number of cases being juggled in the US soars to more than 140."*
>
> *Litigation Trends Survey*[17]

6. Endnotes

[1] Electronic discovery generally refers to the process that organizations use to find, preserve, and produce electronic information that is responsive to lawsuits, investigations, audits, and other formal proceedings.

[2] Fulbright & Jaworski, 2005 Litigation Trends Survey, October 2005.

[3] George Socha and Tom Gelbmann, "The 2005 Socha-Gelbmann Electronic Discovery Survey," August 2005.

[4] Coleman Parent Holdings, Inc. v. Morgan Stanley & Co., Inc., No. CA 03-5045 AI. March 23, 2005.

[5] The Sedona Principles for Electronic Document Production were created by the Sedona Conference® Working Group on Electronic Document production to create "best practices" that would "complement the Federal Rules of Civil Procedure" (and state counterparts) by addressing issues unique to electronic information. Part of the goal of the Principles is to "bring needed predictability and guidance to courts" regarding electronic document production.

[6] Comment 11.a – Search Methodology, The Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production, July 2005.

[7] Fulbright & Jaworski, 2005 Litigation Trends Survey, October 2005.

[8] Medtronic Sofamor Danek, Inc. v. Michelson, 2003 WL 22002514 (W.D. Tenn. August 7, 2003).

[9] "The Surging Demands of e-Discovery," The National Law Journal

[10] Mark Levitt, Robert P. Mahowald, "Worldwide Email Usage 2004-2008 Forecast: Spam Today, Other Content Tomorrow," International Data Corporation, August 2004.

[11] Based on the calculation that the average hard drive installed in a computer sold in 2004 is 80 gigabytes (International Data Corporation), and 1 exabyte is 1 billion gigabytes.

[12] Based on this calculation: 1 typewritten page is equivalent to 2 kilobytes of electronic information (UC Berkeley, "How Much Information 2003") and the average hard drive installed in a computer sold in 2004 had 80 gigabytes (International Data Corporation), or 80, 000, 000 kilobytes, of storage capacity.

[13] Danis v. USN Communications, Inc. 2000 WL 1694325, N.D.Ill., 2000.

[14] Comment 11.c – Consistency of Manual and Automated Collection Procedures, "The Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production," July 2005.

[15] Federal Rules of Civil Procedure, Rule 26(f)(3), as amended and approved by the Supreme Court April 12, 2006.

[16] Stephanie Sabatini, "The Dilemma of Duplicates," Law Technology News, January 15, 2004.

[17] Fulbright & Jaworski, 2005 Litigation Trends Survey, October 2005.

WHERE LAW & TECHNOLOGY MEET

KAHN
CONSULTING INC.

7. About Kahn Consulting

Kahn Consulting, Inc. (KCI) is a consulting firm specializing in the legal, compliance, and policy issues of information technology and information lifecycle management. Through a range of services including information and records management program development; electronic records and email policy development; Information Management Compliance audits; product assessments; legal and compliance research; and education and training, KCI helps its clients address today's critical issues in an ever-changing regulatory and technological environment. Based in Chicago, KCI provides its services to Fortune 500 companies and government agencies in North America and around the world. Kahn has advised a wide range of clients, including International Paper, Dole Foods, Sun Life Financial, Time Warner Cable, Kodak, McDonalds Corp., Hewlett-Packard, United Health Group, the Federal Reserve Banks, Ameritech/SBC Communications, Prudential Financial, Motorola, Altria Group, Starbucks, Mutual of Omaha, EMC Corp., Merck and Co., Sony Corporation, Microsoft, and the Environmental Protection Agency. More information about KCI, its services and its clients can be found online at: www.KahnConsultingInc.com

WHERE LAW & TECHNOLOGY MEET

KAHN

CONSULTING INC.