

accounting for
WORDS



Text analytics technology may help internal auditors uncover hidden risks and gain greater insight on business performance.

DANIEL TORPEY, CPA
PARTNER
ERNST & YOUNG

VINCENT WALDEN, CFE, CPA
SENIOR MANAGER
ERNST & YOUNG

ILLUSTRATION BY RICHARD TUSCHMAN

tHE ROLE OF INTERNAL AUDITORS TRADITIONALLY has been dominated by financial reporting, special projects, and compliance-related efforts. But auditors now are constantly challenged to contribute to overall business performance in a tangible way. In the current environment, practitioners must have access to information to help identify the many categories of enterprise risk — wherever possible preempting the incidence of certain risks and mitigating the effects of others.

Most computer-assisted internal audit tests focus on the numeric data contained within structured sources, such as financial systems and transactional databases. But according to Gartner Research’s “Introducing the High-performance Workplace: Improving Competitive Advantage and Employee Impact,” unstructured or “text based” data, such as e-mail, documents, and Web-based content, represents an estimated 80 percent of enterprise data within an organization. When assessing written communications or correspondence about a key business event, internal auditors often are limited to reading large volumes of data, with few



automated tools to help synthesize, summarize, and cluster key information points to aid in decision making.

To address the full spectrum of data sources surrounding enterprise risk more efficiently, internal auditors can now incorporate unstructured data or “text analytics” tools into their work plans. Text analytics describes a set of analytical tools that identify, classify, and parse words and clusters of words in electronic documents. The software provides for linguistic searches; recognizes and isolates lexical patterns; and provides additional functionality for extracting words by category, theme, or meaning. Moreover, it enables users to tag and structure search results, interpret the data through use of visual tools, and use predictive techniques. Text analytics also describes the process internal auditors and other professionals use to apply these techniques to solve business problems, independently or in conjunction with query and analysis of fielded, numeric structured data.

Text analytics can provide insight into how business risks are emerging. It can also add to internal auditors’ understanding of the people, transactions, and dates associated with significant events — including the development and incidence of fraud — without having to read hundreds of e-mails, documents, or presentations. The software can help practitioners increase their audit efficiency, gain greater and more meaningful information about business performance and enterprise risk, and support the organization’s compliance efforts. Text analytics tools can be used in the context of a risk-based internal audit, as part of a forensic review of controls or business practices, or during an actual investigation.

RISK ASSESSMENT AND ANALYSIS

Text analytics is a relatively new concept. The software stems from a combination of developments in the fields of litigation support and electronic discovery, counterterrorism and surveillance technology, customer relationship management, and research into the life sciences — specifically, artificial intelligence. The application of text analytics in data review and investigations dates back to the mid 1990s.

Text-mining tools broadly referred to as text analytics help users to extract, group, tag, and analyze associations among identified entities and concepts (e.g., noun

themes) and identify the documents that contain them. They create categories, or hierarchical knowledge representations, to auto-classify documents and extracted data. Furthermore, the tools apply statistical techniques to cluster documents according to discovered characteristics.

Concept-based analysis goes beyond traditional search technology by enabling users to group documents according to a statistical inference about the co-occurrence of similar words.

Text analytics generally is used to examine three main elements of target data: the “who,” “what,” and “when.” Internal auditors incorporating analytics into their existing numeric tests would typically use the tools along these three areas.

THE WHO: SOCIAL NETWORK ANALYSIS According to a study conducted by the research firm Meta Group Inc., now owned by Gartner Inc., 80 percent of business people surveyed prefer using e-mail to using the telephone. Most business transactions or events, then, likely have e-mail communication associated with them. Unlike telephone messages, e-mail contains rich metadata — information stored about the data, such as its author, origin, version, and date accessed — and can be documented easily. For example, to monitor who is communicating with whom in the purchasing department, and conceivably to identify whether any relationships therein implied might signal anomalous activity, an internal auditor might wish to analyze metadata in the “to,” “from,” “cc,” or “bcc” fields in department e-mails.

Many technologies for parsing e-mail with text analytics capabilities are available on the market today, some stemming from investigations and electronic discovery software. These technologies are similar to social network diagrams used in law enforcement or in counterterrorism efforts. They enable users to dynamically map communications between individuals, as demonstrated in the “Who” section of “The Three Elements of Text Analytics”

on page 43. Internal auditors should keep in mind, however, that some countries may limit the organization’s access to e-mail data.

THE WHAT: CONCEPT MAPPING The ambiguity inherent in human language presents

significant challenges to the internal auditor or forensic investigator trying to understand the circumstances and actions around an event. This difficulty is compounded by the tendency of people within organizations to invent their own words or communicate in code.

Language ambiguity can be illustrated by examining the word *shred*. A simple keyword search on the word might return not only documents that contain text about shredding a document, but also those where two sports fans are having a conversation about “shredding the defense,” or even e-mails between spouses about eating “shredded chicken” for dinner. Hence, e-mail research analytics seeks to group similar documents according to their semantic context so that documents about shredding as concealment or covering up an action would be grouped separately from casual e-mails about sports or dinner — thus markedly reducing the volume of e-mail requiring more thorough review.

Concept-based analysis goes beyond traditional search technology by enabling users to group documents according to a statistical inference about the co-occurrence of similar words. In effect, text analytics software allows documents to “describe themselves” and group themselves by context, as in the “shred” example. Because text analytics examines document sets and identifies relationships between documents according to their context, it can produce far more relevant results than traditional keyword searches.

Using text analytics before filtering with keywords can be a powerful strategy for quickly understanding the content of a large corpus of unstructured, text-based data, and for determining what is relevant to the search (see the “What” section of “The Three Elements of Text Analytics”). After viewing concepts at a high level, subsequent keyword selection becomes more effective by enabling users to better understand the possible code words or company-specific jargon. They can develop the keywords based on actual content, instead of guessing relevant terms, words, or phrases up front.

THE WHEN: DOCUMENT THREADS In striving to understand the time frames in which key events took place, auditors often need to not only identify the chronological order of documents (e.g., sorted by or limited to dates), but also link related communication threads, such as e-mails, so that similar threads and communications can be identified and plotted over time. A thread comprises a set of messages connected by various relationships; each message consists of either a first message or a reply to or forwarding of some other message in the set. Messages within a thread are connected by relationships that identify important events, such as a reply vs. a forward, or changes in correspondents.

Quite often, e-mails accumulate long threads with similar subject headings, authors, and message content over time. These threads ultimately may lead to a decision, such as approval to proceed with a project. The approval may be critical to understanding business events that lead up to a particular journal entry. Seeing those threads mapped over time can be a powerful tool when trying to understand the business logic of a complex financial transaction.

In the context of fraud risk, text analytics can be particularly effective when threads and keyword hits are examined in light of the Fraud Triangle. Developed in the 1950s by criminologist Donald Cressey, the Fraud Triangle attempts to explain why people commit fraud. Cressey’s premise was that all three components — incentive/pressure, opportunity, and rationalization — are present when fraud exists. The “When” section of “The Three Elements of Text Analytics” illustrates an analysis of the keyword frequency based on the Fraud Triangle.

This analysis method can be applied in a variety of business contexts where increases in the frequency of certain keywords — related to incentive/pressure, opportunity, and rationalization — can indicate risk.

SPOTTING COMPLIANCE RISK

During a company’s internal risk assessment, compliance risks typically can be mapped to divisions or departments using text analytics. These units usually comprise people, and people in today’s organizations typically generate large amounts of e-mail, documents, and other forms of unstructured data.

A large pharmaceutical company’s experience with text analytics illustrates how

the technology is applied in practice. The U.S. Food & Drug Administration (FDA) suspected the company’s salespeople were inappropriately pushing products to certain doctors, making “subtle references” to off-label capabilities. The chief auditor knew his company had already addressed compliance risk in the sales group years ago, and several analytical mechanisms were already in place to monitor such activity. However, data analysis tools couldn’t pick up the subtleties of language and implication in everyday communications and sales presentations, and the chief auditor knew he couldn’t ignore the risk.

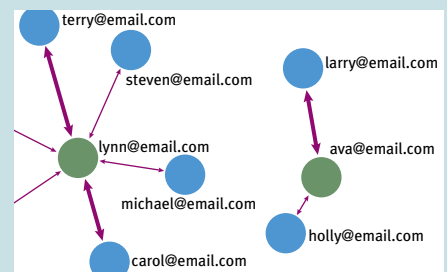
To address the FDA’s concerns, the auditor needed a way to find specific words and phrases used in proximity, such

The Three Elements of Text Analytics

WHO: Social Networking
(E-mail Link Analysis)

Who is talking to whom?

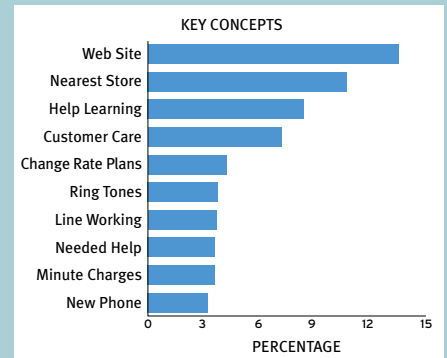
- People to people analysis.
- Entity to entity analysis.
- Mapping of communication lines to organization chart.
- E-mail or phone records, etc.



WHAT: Language Concept Clustering
(Natural Language Processing)

... about what?

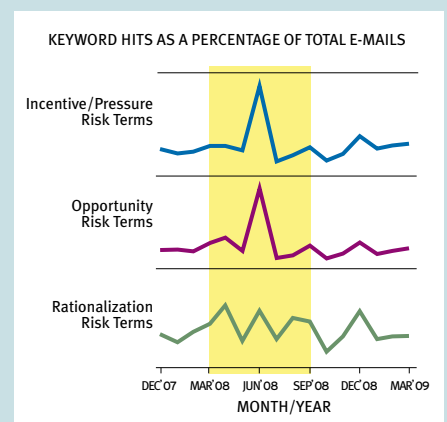
- Top words mentioned in text.
- Key concepts/topics.
- High or unusual dollar amounts.
- Sensitive words/phrases.
- Social Security numbers.
- Credit card numbers and other personal, private information.



WHEN: Communications Over Time
(Fraud Triangle Analytics)

... over which time period?

- When key communications occur.
- In the context of the Fraud Triangle or specific business risk.
- Communication spikes around key business events.
- Communications related to a certain high-risk topic between certain high-risk people.



as e-mails mentioning a specific product along with implicit claims about the drug's efficacy in uses for different conditions. Using a text analytics approach, he was able to leverage text-mining technology to analyze the large volume of data associated with the sales force. Text analytics software enabled him to identify frequent concepts or phrases that may have been used by the group as code words. The technology helped the auditor isolate key nouns, or noun phrases, within the data such as dollar amounts, locations, doctors' names, and drug identification codes mentioned in the text. The chief auditor and his team were able to quickly analyze the large volume of data and narrow the scope of review to a manageable volume of data that showed high-risk keyword hit frequencies spiking within communications to specific customers. The analysis results warranted additional interviews and analysis, which helped focus the investigation on specific people and departments, and ultimately helped to reduce the volume of irrelevant documents.

PROACTIVE INVESTIGATIONS

Instead of relying on electronic discovery software only for data analysis after a fraud or significant adverse event, organizations can consider using the technology to screen data for potential issues before they become unmanageable. Text analytics tools can be used proactively within an enterprise to understand risks and identify anomalies.

At a global technology firm, for example, an internal audit director used text analytics tools to assess compliance risk and help prevent regulatory violations. The firm's audit committee was alarmed by a recent increase in regulatory enforcement activity associated with the U.S. Foreign Corrupt Practices Act (FCPA) and wanted to better understand the risk and dynamics of three recent acquisitions. In less than two months, and with minimal staff and budget, the organization's audit director needed to determine the level of exposure to FCPA violations and other compliance issues. He had to devise a method for examining operations, scanning activities at each of the companies acquired, assessing risk, and providing a meaningful report to the committee.

The basis for the risk assessment — the actions, behavior, and communications of key personnel in disparate operations —

comprised not only books and records, but e-mail and other documentation. To sort through this data, the director incorporated text analytics into an FCPA risk assessment framework, enabling him to evaluate certain enterprise directory folders in the sales and acquisition departments and to test critical areas. Data sources included sales presentations, Microsoft Word documents, PDF files, and key e-mail communications of individuals selected through the use of text analytics. The director's audit team searched for patterns related to the people, events, transactions, and dates associated with key business communications surrounding the acquisitions in question. Ultimately, he and his team were able to assess risk exposure quickly and report back to the audit committee, without needing to read every document.

LIMITATIONS

In his *Summa de Arithmetica* of 1494, the father of modern accounting, Luca Pacioli, noted, "Though a businessman's head have a hundred eyes, still there are not enough for all of his duties." Text analytics provides an extra set of eyes, but even with increased vision, users must use this tool with care and consider its practical limits.

DON'T BOIL THE OCEAN Considering the overwhelming amount of text-based data within today's enterprise, internal auditors could never hope to analyze all of it; nor should they. The exercise would prove expensive and provide little value. Just as an auditor would not reprocess or validate every sales transaction in a sales journal, he or she would not need to look at every e-mail from every employee. Instead, the auditor would take a risk-based approach, identifying areas to test based on a sample of data or on an enterprise risk assessment. For text analytics work, the auditor may choose data from five or 10 individuals to sample from a high-risk department or a newly acquired business unit.

DON'T EXPECT GUARANTEES No matter how sophisticated the search and information retrieval tools used, there is no guarantee that all relevant or high-risk documents will be identified in large data collections. Moreover, different search methods may produce differing results, subject to a measure of statistical variation inherent in probability searches of any type. Just as a statistical sample of accounts receivable

or accounts payable in the general ledger may not identify fraud, analytics reviews are similarly limited.

INCORPORATE, BUT DO NOT REPLACE Text analytics can be powerful when integrated with traditional internal audit data-gathering and analysis techniques such as interviews, independent research, and existing audit tests involving structured, transactional data. For example, an anomaly identified in the general ledger related to the purchase of certain capital assets may prompt the internal auditor to review e-mail communication traffic among the key individuals involved, providing context around the circumstances and timing of events before the entry date. Furthermore, the internal auditor may conduct interviews or perform additional independent research that may support or conflict with his or her hypothesis. Integrating all three of these components to gain a complete picture of the business event can yield valuable information. While text analytics should never replace the traditional rules-based analysis techniques that focus on the organization's financial accounting systems, it is also important to consider the communications surrounding key events typically found in unstructured data, as opposed to the financial systems.

ENHANCING PERFORMANCE

Internal auditors increasingly are being challenged to help improve overall business performance. Last year, a *Wall Street Journal* article focusing on text analytics mentioned that "a growing number of corporations are now discovering that text analytics can help them identify market trends and customer patterns, spot fraud and security threats, and highlight problems with products."

Traditionally, stakeholders have used internal audit functions to keep their companies out of trouble. By incorporating internal audit methodologies around text analytics, auditors can also enhance their proactive risk efforts and potentially improve business performance for the clients they serve.

To comment on this article, e-mail the authors at daniel.torpey@theiia.org.

For additional examples of text analytics applications, visit InternalAuditorOnline.org.