©2013 ARMA International • www.arma.org

Rolling the Dice with Predictive Coding

Leveraging Analytics Technology for Information Governance

Leigh Issacs

©2013 ARMA International • www.arma.org

With increasing frequency, organizations are realizing the importance, value, and benefits of implementing an information governance program. Information continues to proliferate at an insanely rapid pace, and the combination of legacy unmanaged information and the plethora of new information that is being generated in mountainous volumes becomes overwhelming.

New types of vehicles and repositories that deliver and store information are also increasing exponentially. Structured and unstructured data are found stored in archives, e-mails, hard drives, and other data repositories, resulting in significant pockets of information that are difficult to access and search. The financial, productivity, and time costs of maintaining storage systems for so much data are prohibitive.

To complicate matters, it is difficult to gain a clear picture of what this data contains, posing significant business risks and adding to the time and expense of discovery or investigation exercises. As such, it is imperative that organizations act on better identifying and reducing their volumes of data.

Information governance is a strategic framework composed of standards, processes, roles, and metrics that hold organizations and individuals accountable to create, organize, secure, maintain, use, and dispose of information in ways that align with and contribute to the organization's goals.

Source: *Glossary of Records and Information Management Terms, 4th Ed.* (ARMA TR 22-2012).

The data growth projections and statistics are staggering. Recent estimates by analysts and research indicate that:

- Each year 1,200 exabytes of new data will be generated, according to the IDC Report "Business Strategy: Business Analytics and Big Data – Driving Government Businesses."
- Enterprises will experience 650% data growth in the next five years, says David Rosenbaum in his report "That New Big Data Magic" on *cfo.com*.
- 80% of this data will be unstructured and generated from a variety of sources, such as blogs, web content, and e-mail, states Adrian Bridgewater in a report on CWDN: The Computer Weekly Application Developer Network.
- 70% of this data is stale after 90 days, notes Big Data Bytes in "90% of Everything Is Crud."

At a glance, these statistics make it clear that information governance challenges will continue to grow as organizations attempt to separate the wheat from the chaff to determine what is useful versus what is not and what information exists beyond its usefulness or identified retention requirement.

At the same time organizations are struggling to find the magic wand to conquer their information governance woes, technologies are emerging to address large volumes of electronically stored information (ESI) for purposes of e-discovery. Organizations should consider whether it is possible to leverage these tools for identifying, managing, and appropriately governing their information.

Technologies for Tackling IG Challenges

When used innovatively, the solutions used during e-discovery and litigation to avoid cost and time to address legal hold, data collection, data processing, and document review can also be expanded to add value into other areas, such as information storage, records retention, and data security.

Predictive coding, which is an evolving technology that combines people, technology, and workflows to find key documents and identify and review large data sets, is one of those solutions. More specifically, predictive coding is machine-learning technology that "trains" the computer program to automatically predict how documents should be classified, based on initial human guidance. The computer program then applies what it has learned to the universe of information.

Predictive coding can be used in three ways to actively address the classification issues associated with current information: data remediation, classification of information already in repositories, and classification of information upon its creation.

For data that can't be classified automatically, technologies can be used (e.g., sampling and manual review) to make content-driven decisions that can then lead to defensible retention and disposition. (See Doug Smith's RIM Fundamentals article, "Thinking Outside the Box: Use Predictive Coding as a RIM Tool," on page 30 of this

©2013 ARMA International • www.arma.org

issue for a primer on how these technologies work.)

The combination of these technologies and processes allows organizations to:

- Highlight data that may present a business risk
- Find, retain, and profit from valuable data
- Remove redundant, irrelevant, or stale data

In February 2012, Hon. Andrew J. Peck of the U.S. District Court for the Southern District of New York entered an order authorizing the parties in *Da Silva Moore v. Publicis Groupe, et al.* to rely on predictive coding for identification of responsive documents during discovery in lieu of traditional document review and search terms.

Peck's decision cited statistics indicating human review is no more accurate, and perhaps less accurate, than computer-assisted review and lends support to the accuracy of these technologies.

In fact, according to a blog posted by Maura R. Grossman, Esq., by leveraging such things as document type, language, content, party, timeframe, individual name, and conceptual meaning, predictive coding has been known to generally deliver more than a 95% accuracy rate.

Predictive coding and analytics can search and manage the abyss of existing archives to identify data that should be kept or deleted . . .

When considering the capabilities of predictive coding technology, it is not a stretch to connect its usefulness in e-discovery with the potential to add benefits to a proactive IG program. It is difficult and time-consuming to sift through e-mails, texts, contracts, spreadsheets, and other types of ever-increasing media. While subject matter experts are still necessary, predictive coding eliminates much of the manual work and time from the task.

Areas Where Predictive Coding Can Help

Predictive coding technologies allow for the identification, review, and tagging of information – and can frequently apply and execute specific workflows. Thus, a few areas of an organization's digital landfills that can benefit from the use of predictive coding and analytics tools are listed below.

Retention, Disposition

Organizations have an obligation to manage information appropriately. Many industries are highly regulated and subject to other legal or legislative recordkeeping requirements. Utilizing subject matter experts paired with predictive coding technologies increase the accurate identification of information and thus provide a sound foundation for defensible disposition and prevent over-retention of expired or unnecessary content.

Archives

Believing that archiving was the panacea of information management and would solve all of their data storage woes, many organizations jumped on the archiving bandwagon. In reality, the archive resulted in the equivalent of throwing hundreds of thousands of loose papers in a room and shutting the door. Content was not identified and organized in a manner that allowed it to be governed and managed, and archives were not designed to be searchable in any meaningful way, resulting in a mess.

Predictive coding and analytics technologies can search and manage the abyss of existing archives to identify data that should be kept or deleted, and often minimize the costs and risks of potentially moving unnecessary data. These technologies can span the archives and locate data that:

- May be subject to litigation hold
- Is important intellectual property or a vital record
- May contain sensitive information
- Can be defensively deleted, such as duplicate data

Legal Holds, Protective Orders

It can't be preserved if it can't be found. It will be over-retained if an organization can't figure out what it is. Holds can't be lifted if an organization doesn't know what it has, and retention cannot be reinstated once the hold is lifted. Predictive coding solutions can add value long before and long after the collection process.

Data Remediation

How often has an organization found important, business-critical information that has been misfiled? Predictive coding solutions can provide a tool to locate various types of vital records and contracts, and, if necessary, move them to other repositories where they can be secured and managed. It can also delete redundant, obsolete, and trivial files, which, as learned earlier, can consist of up to 70% of the organization's data.

File Transfers

Some organizations (e.g., law firms) have accepted information mobility and lateral movement as a routine course of business. Ethical requirements often dictate what types of, to whom, and when information can be re-

©2013 ARMA International • www.arma.org

leased. The ability to quickly and efficiently identify, segregate, and filter various types of electronic information is critical. Predictive coding and analytics solutions can assist in reducing the manual efforts historically performed by higher value timekeepers that could be better utilized on billable activities.

Intellectual Property, Knowledge Management

Appropriately identifying, maintaining, and securing an organization's intellectual property can be the difference in its success or failure. Using these technologies, it is possible to cull through data to identify potential pockets of information that should be retained as valuable intellectual property, and thus, protect the organization's vital assets. Many organizations also rely heavily on their bank of information that serves to support and pass along various pockets of knowledge and templates. To do this effectively, an organization must have efficient tools with which to identify and segregate this information from other classifications of data and, in some cases, move it to another repository.

Business Intelligence

Analytics technologies can reveal the value (or lack of value) of specific unstructured data by mining it for business insights and intelligence. Once the value has been identified, it can then be leveraged in various ways to benefit the organization, including strategic planning and decision making.

Sensitive Information

Organizations continue to see growing regulations and requirements for the management and security of sensitive information. Sensitive information may be created and stored across the enterprise, often in unstructured systems. Analytics technologies can aide in the identification of this type of content to better support compliance.

Data Security, Fraud

Data security and fraud continue to be a concern. Finding clues to these issues amongst the vast amount of information can be daunting. Often, by the time a security breach or fraud has occurred, it is already too late. A proactive approach that provides an opportunity to appropriately deal with offenders can protect the organization and its information.

Using dynamic indexes, workflows, and automated risk analysis and monitoring, these types of technologies can help identify fraud and other breaches. These technologies can also provide the ability to flag risky content for review and escalate potential risks appropriately within the organization. Having such programs in place can help lower an organization's insurance premiums as it has taken measures to proactively identify potential risk.

File Shares

There are various strategies to attack the monster of uncontrolled file shares that are notorious for containing masses of poorly classified, and sometimes irrelevant, data. Even if processes exist to address the information lifecycle of file share content, organizations typically lack the tools to defensively delete the information after it is no longer useful or required to be kept.

Predictive coding tools can help classify and manage the data in the file shares to help remediate these challenges.

The sheer nature of file shares promotes the proliferation of duplicative content or unnecessary data. However, in the haystack of this unnecessary information may be the needle – pockets of valuable information that need to be kept or information that needs to be securely maintained and managed. Predictive coding tools can help classify and manage the data in the file shares to help remediate these challenges.

The Next RIM Tool?

As many records managers will attest, it's often difficult to get buy-in and budget approval to acquire the tools helpful and necessary to automate and support IG and accomplish the tasks outlined above. However, given the recent e-discovery boon and the footprint that predictive coding and analytics technologies have already developed in that area, there is an increasing trend for organizations to have these types of solutions in-house.

It is no longer effective to govern information by leading with a "big stick" and setting strict rules and regulations for individuals to follow. Risk and compliance will always be a key requirement of an efficient IG program. However, an organization cannot lose sight of the fact that to be effective, its IG program must support the business rather than be a hindrance to it.

Smarter, more efficient approaches to coding and classifying information – such as by using predictive coding technologies – not only help organizations manage risk and achieve compliance, they also allow them to operate more effectively and efficiently.

©2013 ARMA International • www.arma.org

When individuals are able to do their work without worrying unnecessarily about overly burdensome retention schedules or belaboring where information should be filed – and are still able to find the information when they need it – the organization has added a layer of operational efficiency while at the same time governing its information assets.

After having seen practical ways in which predictive coding technology can be an asset for routine functions performed on information throughout every phase of its life cycle, it begs consideration that it may emerge as the next tool in the records professional's toolkit.

These technologies are here to stay and, while disputes about their uses and limitations from a litigation and e-discovery standpoint will continue to be debated in our courts, there is value to exploring them for real, practical, tangible tasks in governing information throughout its life cycle. **END**

Leigh Isaacs can be contacted at leigh_isaacs05@yahoo.com. *See her bio on page 46.*