RISE OF THE MACHINES

THE ROLE OF TEXT ANALYTICS IN RECORD CLASSIFICATION AND DISPOSITION

James Santangelo

The disparate nature of unmanaged data repositories hinders an organization's ability to purge unneeded data because the size and nature of the repositories make it cost-prohibitive to evaluate the data for disposition.

The appropriate information management strategy, along with recent developments in text analytics search technology, can be combined to help solve the problem of large, unmanaged data repositories that increase an organization's cost and legal exposure.

However, if shared file data repositories continue to be left unmanaged, organizations will have vast wastelands of unstructured electronic information. They will be unable to distinguish among data that does and does not need to be retained, and they will be locked into the ever-increasing cost of storing and maintaining the data in order to comply with their legal obligations.

Accurate classification of electronic information, or identifying and associating information types with electronic data, is essential to making the appropriate retention and disposition decisions. To reduce volume, organizations must be able to determine what type of data they have to understand what data must be retained and what data are no longer useful. And reducing the volume of data is one generally accepted approach to reduce data's primary risk – the high cost of finding, preserving, reviewing, and producing it for litigation.

Accurately classifying data allows organizations to accurately retain it, place legal holds on it, and make reasonable disposition decisions about it, thus helping to minimize the significant legal costs and risks associated with continuing to store it unnecessarily. But, because of the seemingly complex, costly, and insurmountable task of classifying many years worth of unmanaged data, little has been done to address the problem.

Human Neglect

Many organizations rely on their records management policy and retention schedule to provide employees with the guidelines they need to apply retention to their data. However, employees' primary responsibilities may leave them little time to do these administrative tasks. As a result, even trained employees may fail to accurately determine how long a file should be retained, to what record classification it belongs, or how long it must be preserved for litigation.

The approach of relying on policy and assigning employees with the sole responsibility of retention, even though the data is actually owned by the organization, can cost the organization a great deal of money. Cohasset Associates' *Information Governance: A Core Requirement for the Global Enterprise* says that regulators and courts may still hold an organization responsible for its employees' actions in this regard.

Classification Ruin

The term classification has different meanings depending on its use. In an information management context, it can be summarized from ISO 15489-1:2001 *Information and Documentation – Records Management – Part 1: General* as a standardized way of identifying and arranging records into categories according to a logically structured classification scheme. The categories typically correspond to record classes that are indicated in a company's record retention schedule and may contain a number of different, but related, record types.

Each record class delineates specific types of business information or data, with each class being retained for a specific amount of time from days, to years, to indefinite, thus implying the data's eventual destruction once all legal hold obligations are fulfilled. Yet, many times, the destruction never takes place because not enough identifying information is known about the data to make a decision on its disposition.

If data exists as individual electronic files stored in formally managed shared folders according to specific business functions' file plans, its classification can be based on its content and retention can be determined.

Conversely, if the data has not been formally managed, which is the typical scenario of today's average large organization, the shared folders can store millions of different types of electronic files. These files may contain many other classes of records and non-records, files that range in age spanning decades, and many drafts and copies of the original files. In this case, files exist in a vast, disorganized environment where they will be difficult to evaluate, identify, and, hence, to classify.

Lack of data classification not only prevents its disposition, driving up discovery legal risk, but also can have regulatory, security, and knowledge risk implications. Penalties associated with regulatory audits can result from not being able to retrieve data when it is needed, hindering an organization's ability to meet its compliance obligations. It also makes an organization more vulnerable to data breeches or leakage because unorganized data is more difficult to safeguard.

Organization-wide productivity also suffers from the inability to find relevant data. If staff members cannot leverage knowledge, the organization's ability to remain competitive is affected.

Accurate classification of data is a factor in all of these risks and others. For highly regulated companies with significant risk of litigation and with many terabytes of unstructured data, millions of dollars are at risk, as shown in the 2004 case *United States v. Philip Morris USA Inc.*

The organization's policy may state that retention is the responsibility of the employees, but asking employees to spend a large amount of time manually classifying data greatly affects productivity, which translates into a sizable financial impact.

Human vs. Machine

Legal and compliance executives may recognize the increased risk of storing the data, but information must be classified before disposition decisions can be made. The organization's policy may state that retention is the responsibility of the employees, but asking employees to spend a large amount of time manually classifying data greatly affects productivity, which translates into a sizable financial impact. IT executives are often approached for a technology solution to automatically classify the data, but most available technologies are costly to implement.

Aside from the cost issues the manual or automated methods impose, lack of accuracy affects both methods and creates a dilemma about which method to use. On one hand, it is difficult to rely on employees, who may not have the records knowledge or the time to consistently and accurately classify information. On the other hand, there is a lack of trust regarding the accuracy of technology tools that use extensive pre-defined policy rules to classify information. So, if an organization cannot rely on employees or technology tools, how can its information be accurately classified?

The nature of this dilemma exists because organizations are either asking employees to do a job – some of which could be handled better by technology – or technology to do a job, some of which is better handled by employees. Using only the manual method or only the automated method to classify information will leave a company with a mix of information, some of which is classified incorrectly and some of which is left unclassified. Employees and technology must work together, but how?

Humans in Control

When using a manual method for classification, employees still need the technological means with which to store the files. How employees use technology to store files according to their classification can vary, depending on the type of files being stored and the systems used by employees.

Having the folder structure align with the records classification scheme is one way to help employees classify them. Using this method, files are stored within a folder hierarchy that indicates the file's classification, and employees can use their discretion in making retention decisions. E-mail and attachments that need to be retained as records also can be saved in this hierarchy along with other related files in their business context.

However, this method requires employees to periodically revisit their files because the length of retention often depends upon the date of a particular event. For example, the retention period for a contract may not begin until the contract expires. Subsequently, this manual method requires employees to thoroughly and frequently review their files to determine whether they are eligible for destruction. This typically does not happen because of employees' time constraints or other priorities.

The disposition review task would then be left to a large staff of IT or records administrators, who would need to work with the employees on a regular basis to make the retention determinations, which is also a costly alternative. This leads organizations to consider whether using a records management application would help reduce the effort.

When official records need to be stored, such as for regulatory purposes, a records management application may be beneficial. It facilitates classification, allows the application of other important tags, and provides other records management functions, including control of disposition.

Without a regulatory need, however, a large-scale deployment of a records management application is challenging to implement due to the unusual amount of manual employee effort it can take to store files in it and then maintain it. For example, it requires extra effort for an employee to enter a multitude of metadata fields to store a document. When the effort threshold is not justified, or if employees find it difficult to simply store a file, they will not use the system, bringing the organization back to square one.

Machines in Control

To alleviate the burden on employees presented by manual classification, organizations can use varying degrees of automated classification tools to accomplish the task. Automated tools can evaluate, identify, and classify large quantities of files and can be used to help support employees' efforts.

For example, automated tools can be used to classify and then organize existing shared folder repositories and archives, or they can work in the background to classify files as employees store them in a repository. The tools take the responsibility of classifying the files through terms analysis and can also extract from the text of the files metadata useful for providing the additional information needed for determining retention.

Variations in automation levels of available tools to accomplish this can span from highly automated, where the tools perform all the work, to semi-automated, where the tools are used to provide suggested classification tags to the employees as documents are being stored.

The implementation of highly automated classification tools require the development of policy rules that stipulate how the files should be classified and retained. Typically referred to as policy rule engines, these tools use the word index of each electronic file, along with predefined classification rules, to match the information in the file to its record classification.

The rules can be built to identify information based on terms found within each electronic file. For example, with employment records, these tools can index the content of a file and then use terms, such as "performance evaluation," in its analysis to categorize the file as information pertaining to the employment record classification.

However, unless other variables are incorporated as a cross-reference, this automated analysis may erroneously apply this classification to an unrelated subject with similar terms, such as performance evaluations of mechanical engines. Considering the multitude of varying types of information organizations generate on an enterprise-wide basis, the magnitude of the challenges in implementing and maintaining these tools is evident. The accuracy of these automated tools is dependent on the accuracy of an ever-changing, complex set of policy rules.

The policy rules must first be set appropriately to promote the accurate processing of files. In non-technical language, the logic of a rule may state: "If the file contains the terms 'performance,' 'appraisal,' 'employee,' 'review period,' and 'rating,' then classify the document as belonging to the performance appraisal classification."

The latest advancements in text analytics use sophisticated techniques to determine the conceptual meanings within each file to compensate for shortcomings and extend the functionality of the applications that use policy rule engines.

The system can then be set to determine the employee ID and the year of the appraisal and pull that information from the metadata of each file for the purpose of retention. However, when creating the rules, many test iterations need to be performed, either initially for each record type, or later if changes to the information occur. This is compounded by having many different repositories or many different systems.

So, in this case, if a modification to a key term, such as the employee ID, is made, the rules would need to be reviewed and updated to not affect how the data is classified. Given these complexities, detailed policy rules are generally impractical to maintain on a large scale.

Rise of the Machines

The latest advancements in text analytics use sophisticated techniques to determine the conceptual meanings within each file to compensate for shortcomings and extend the functionality of the applications that use policy rule engines. Use of text analytics greatly increases the accuracy of the classification by interpreting the meaning of terms in their context instead of being limited by the character strings inherent in policy rule engines.

One such method, called natural language processing (NLP), determines the contextual meanings of terms by relating them to other terms in the file incorporating thesauri (synonyms), stemming (variations of the same word), and other analysis in the process.

A common example that illustrates the use of text analytics to evaluate terms in a file is how the system distinguishes the word "sue" from either referring to a person's name or to a legal action. Its ability to reliably and accurately classify files has increased; however, building and maintaining the rule sets are still a concern with this form of automated system.

Certain types of text analytics technology that make use of Bayesian Inference, for instance Adaptive Probabilistic Concept Modeling (APCM) or Probabilistic Latent Semantic Analysis (PLSA), form the basis of certain products and can provide accuracy with minimal rule building and maintenance, according to *Document Categorization Using Latent Semantic*

Indexing and Linear Prediction Models with Graph Regularization for Webpage Categorization.

These methods use patterns combined with mathematical calculations, instead of character strings, to attempt to best determine how to classify a document. The successful use of this system is based on its ability to "learn" from existing sample files in order to create and maintain these patterns that correspond to the company's classification scheme.

... if files are stored within a content management system with sufficient records management capabilities, holds and retention can be applied by the system automatically, allowing for automated disposition.

Multiple samples of each file type, including representative variations, are analyzed by the system to help it determine the typical identifying characteristics of specific file classifications. It then uses complex mathematical calculations to process the data and cluster each file as it relates to one or more concepts. This method also may help find unknown relational concepts that cannot be easily determined by people.

Additionally, certain systems adjust continually as they analyze files to maintain their relevancy as the content of the files changes over time. This method significantly increases the accuracy of the process, which can be as high as 90%, according to *Linear Prediction Models with Graph Regularization for Webpage Categorization and Topic-bridged PLSA for Cross-Domain Text Classification.* It also greatly minimizes the effort needed to initially set up and maintain the system. And although the technology used alone cannot pull metadata or determine retention events, it can be used to provide a reasonable foundation to tackle the problem.

Certain vendors that provide classification applications use APCM or PLSA technology as the basis of their analytical engines. Other vendors combine it with other complementary technologies, such as Boolean and NLP, to provide increased capabilities to help handle metadata. These tools can be a good solution for legal, records, and IT professionals who need to clean up their organization's existing data stores, but that will ultimately require employee input to validate the results.

Using this approach will greatly reduce the effort needed to sort through masses of data, opening the door for organizations to attempt the task. It also will set the stage for implementing a controlled environment for sharing files that will help prevent the situation from reoccurring.

Humans and Machines Coexist

Depending on how the files are stored and used, a combination of manual and automated classification methods can provide good results. While these methods and tools provide the means to automate the disposition process of most unstructured data, files where the retention event is variable or where legal holds must be applied also must be accounted for.

Overcoming these challenges requires implementing a content management environment where information is managed in an organizational hierarchy that is aligned with a retention classification based on business functions. The system also must have the ability to apply retention and legal holds to the data it contains and provide a flexible means to determine or set retention event dates.

Most content management systems can automatically determine retention where the system uses classification along with a file date, or the typical life span of that type of information, to estimate the destruction date. In other cases, these systems have the ability to prompt or remind employees to enter the retention event date. Either way, if files are stored within a content management system with sufficient records management capabilities, holds and retention can be applied by the system automatically, allowing for automated disposition.

Conclusion

The strategy needed to have retention and disposition applied to unstructured data stores consists of the initial cleanup of data and its ongoing maintenance. The initial cleanup involves a combination of automatic and manual classification where the appropriate technology is used to process the files, which are then reviewed by knowledgeable employees to verify accuracy.

Ongoing maintenance involves the use of content management tools with records management and legal hold capabilities that include processes to handle different retention events. The methods discussed in this article, when implemented in the right combination, offer an effective approach that can provide executives and administrators with the information they need about their organization's data in order to make current and ongoing disposition decisions confidently.

James Santangelo can be contacted at james.santangelo@us.pwc.com. *See his bio on page 47.*