

# The Power Of Automated Categorization Technology: How To Handle Information That Is Here Today, Here Tomorrow

Jason Robman

RECOMMIND INC.

Over the last five years, the distinction between records and knowledge management, information access, legal and compliance systems has blurred. Whereas each was once a separate IT system, they are no longer the sole domain of one department as the business drivers behind each system become more and more interconnected. The primary unifying factor across these systems is that they manage corporate information in electronic form; yet while each system uses enterprise data for a different purpose, the data itself is largely the same. And unless the data is organized, it becomes exceedingly difficult to use and costly and inefficient to manage. This growing need for the organization of ever-growing enterprise information has led to skyrocketing interest in automated categorization technology.

In basic terms, automated categorization technology classifies large volumes of data with much greater speed and accuracy than humans alone. Automated categorization can now accurately classify millions of newly-created documents every day, in the process greatly improving the speed and the quality of information management.

## Information Governance As The Driver

With all companies experiencing growing volumes of electronic content, information governance has become a strategic imperative for enterprises. For most enterprises, information is the organization's life-blood, but it can often act more like a raging river when left uncontrolled and unmanaged. The same information that can fuel a company's productivity, success and ability to act quickly can also pose substantial risks when the company is faced with litigation or regulatory compliance

*Jason Robman is Director of Legal Solutions and Corporate Counsel at Recommend Inc., the leader in search-powered information risk management (IRM) software. Jason helps Recommend's enterprise customers effectively manage their regulatory and eDiscovery risk, while also handling corporate counsel duties for the company. A member of the California Bar, Jason is active in the Sedona Conference and is also a co-chair of the EDRM Project's Search Working Group. Prior to joining Recommend, Jason was with FTI Consulting where he focused on complex litigation matters, securities litigation, antitrust matters, internal corporate investigations, and electronic data analysis and computer forensics engagements. He also assisted clients involved in antitrust and merger issues, intellectual property theft, securities investigations, FCPA investigations, MTBE litigation, and other types of class action litigation.*

*Jason's experience includes several years of conducting internal investigations while employed at Bank of America in its Regulatory Investigations Group. In this role, Jason supervised large scale investigations involving the global corporate and investment bank activities in equity and debt trading, prime brokerage (hedge funds), mergers and acquisitions, syndicated finance, equity research, public finance, private equity, broker-dealer services, and retail wealth management.*

events.

Because people have long been viewed as the most important asset of a company, tens of billions of dollars are spent each year on technology and processes to support human resources, recruiting, training, development and retention of employees. In an information-driven economy, however, one could argue that it is actually enterprise information that is a company's most important business asset. In today's hyper-fast business environment and with the increase in regulatory and legal risks, skyrocketing costs associated with eDiscovery and rising data storage requirements, the goal of better organized information is no longer an option, but rather has become a necessity.

Many companies are now embracing the concept of information governance and even creating a senior level position for an Information Governance and Risk Officer; or, at the very least charging the Chief Information Officer (CIO) with the responsibility of implementing and maintaining sound information governance. The key to information governance is to implement an automated information management system that will meet the organization's compliance, eDiscovery, data protection, and knowledge/records management needs.

## Addressing The 'Legacy Data' Problem

Companies across all industries are faced with bursting stores of electronic data that continue to grow unchecked. Most organizations really don't know what information they have, so they keep everything, creating giant data repositories of unidentified legacy information. As a result, many are faced with outdated systems and inefficient storage that are no longer required for day-to-day business operations, but are being retained because no one knows what to do with it. When coupled with skyrocketing storage costs, the fear of spoliation in a litigation context and regulatory fines for failure to retain (i.e., HR 5811, Sarbanes Oxley, SEC 17(a)-4, etc.), keeping all of this legacy information can paralyze corporate IT departments.

In addition, trying to manage all of this electronic information manually is a virtual impossibility. As many companies have found, relying on employees to organize their information is a losing battle. The only way to keep up with the creation of terabytes of electronic information is to use technology to identify, tag and categorize content based on its meaning. With the right technology, users can get what they want – instant access to relevant data across the enterprise – while the IT, records management and legal departments get what they need, namely an automated, accurate data classification system.

## Challenges Of Unstructured Information

Unstructured information refers to computerized information that either does not have a *data model* or has one that is not easily usable by a computer program, like file share data, wikis and blogs (email is often called "semi-structured" but is generally lumped with unstructured data from an information management perspective). The



Jason Robman

term distinguishes such information from data that is stored in fielded form in databases or annotated (semantically tagged) in documents. Information stored in a computer which is not usable by a computer program is worthless from a business standpoint but continues to take up storage space and IT resources. In order to help users find what they need to perform their jobs, structure (or intelligence) must be added to unstructured data.

There are some clear benefits of categorization technology that address these challenges of unstructured information. First, unlike keyword search engines (i.e., Google, Autonomy, etc.) categorization technology automatically organizes data based on its meaning – irrespective of keywords. Categorization technology allows a single records manager to automatically categorize terabytes of data. The investment return can be seen as technology categorization can save a company millions of dollars per year on direct storage, data center and data archiving costs. Furthermore, a company can make an educated assessment about data retention since the content of the information is now known and understood in a manner that was impossible before applying categorization technology.

## How Categorization Technology Works

Categorization technology crawls and indexes text from any source, including document management systems, intranets, web sites, CRM applications, databases or file systems, and then tags that data to associate it with one or more categories. The categories can be based on the content of the documents themselves or on company-created categories such as topic, document type, geographic location, language or industry. Certain technologies are able to perform automated categorization of new documents based on training, i.e., using positive and negative sample documents as guideposts, while others use user-defined rules. More accurate and sophisticated systems use a combination of both techniques. As documents change and new documents are added within a company's environment, categorization technology can automatically tag and classify them into existing taxonomies and/or newly created taxonomies.

## Key Categorization Technology Features

Key categorization features to be aware of include:

### Automated Metadata Creation

The right categorization technology can make a company's information easier to find by automating the creation and enhancement of metadata, associating documents with categories and annotations on the fly and improving the organization of enterprise information without the need for pre-existing taxonomies.

### Accurate Categorization Based on Content

Companies should be able to dramatically reduce the time and effort needed to organize, tag and manage information. To facilitate these huge improvements, effective categorization technology must use both automated training and rules-based categorization tools to provide content experts with the ability to organize much larger quantities of data than can be handled manually, without sacrificing quality or control.

## Accurate Relevance Reporting

One of the relatively new features of categorization technology is relevance scoring which can show how accurate and inclusive the categorization is. Relevance scores in the form of a confidence percentage can be used to assess the system's confidence in its decisions for each document. In addition, a categorization system can even monitor and track the overall precision and recall rates for each category within the taxonomy that was processed.

In addition, categorization technology should enable a company to:

- Accurately categorize documents into categories in one or multiple taxonomy structures;
- Perform with multiple training modes – automated learn by example, rules-based, or assisted with positive and negative examples;
- Support multiple taxonomies, flat or hierarchical;
- Easily import existing taxonomies;
- Provide quality reporting that identifies accuracy level by category;
- Provide support for documents in multiple language;
- Integrate with myriad enterprise information sources including document management/records management systems, portals, file systems, databases and web sites.

## Adoption

Once the appropriate categorization technology has been identified, a company needs to start thinking about adoption. First and foremost, the objective should be to get the appropriate stakeholders around the same table to identify available operating and capital budgets along with business requirements. Suggested stakeholders include legal (both from a "customer" perspective and to advise on regulatory requirements), records management, information technology, compliance and knowledge management.

Once requirements are established, a subset of the project team should be charged with identifying the best of breed technology solutions that can meet the company's requirements. It should be noted that technology in this area is developing rapidly and will most likely continue to evolve over the next few years. So do your homework by reviewing the offerings thoroughly, speaking with reference customers, and requesting a demonstration with your own data. As with many types of technology projects there is often only one bite at the apple so make sure the project is organized, well scoped, and remains close to budget to facilitate success.

## Conclusion

Automatic categorization technology has truly come a long way in its ability to assist various enterprise constituencies with their daily information management challenges. Professionals from records and knowledge management, information access, legal and compliance should collaborate and be creative on the issue of how their company can benefit from categorization technology as described in this article and beyond. Despite the recession forcing many enterprises to more closely scrutinize funding for data management projects, investment in categorization technology is a powerful counter-weight as the cost savings from reducing the amount of data outweigh the investment many times over.

Please email the author at [jason.robman@recommend.com](mailto:jason.robman@recommend.com) with questions about this article.