Enterprise content management solutions
White paper
April 2009

Information Management software

Achieving compliance and controlling costs with automated categorization of e-mail for records management.

A look at best practices and a U.S. Army use case





Contents

Introduction

E-mail management has become more important in organizations, especially government agencies. The amount of e-mail handled by organizations grows each year, and so does the need to preserve and categorize those e-mail messages for compliance—as well as harness that unstructured information to help organizations work smarter. In June 2008, the Government Accountability Office (GAO) issued report GAO-08-742 on records management. The report, dated June 13, 2008, stated that reported federal agencies needed to strengthen e-mail management. According to the GAO report, agencies need to comply in nine key areas, as defined by the National Archives and Records Administration (NARA) regulations:

- *Agencies must inform staff that e-mail messages are potential records.*
- *Staff must be capable of identifying federal records.*
- *E-mail transmission data and distribution lists must be preserved.*
- *Agencies must state that draft documents circulated on e-mail systems are potential federal records.*
- *E-mail must be stored in an appropriate record-keeping system and staff must be informed of how these records, regardless of format, are maintained in that system.*
- *Agencies must provide instruction on how to copy e-mail identified as federal records from an e-mail system to an official record-keeping system.*
- *E-mail systems must not be used to store record-keeping copies of e-mail messages identified as federal records.*
- *Agencies must not use e-mail backup tapes for record-keeping purposes.*
- *Staff must be educated on the management and preservation of e-mail records sent or received from nongovernmental e-mail systems.*

Highlights

The GAO examined e-mail management policies at four different agencies and found noncompliance in all nine areas.

How can other government agencies avoid the same fate? Simply saving e-mail and other documents isn't enough. Some e-mail messages are federal records that must be maintained, and others are insignificant and can be deleted. Given the volume of e-mail and information that agencies receive, manually categorizing records for easy auditing and more effective compliance becomes less feasible every year. Agencies need to manage records in a way that will help improve efficiency and cost savings and support compliance.

Automated categorization of e-mail can help government agencies stay in compliance, despite an explosion of unstructured content.

How auto-categorization can help agencies conform to regulatory standards

One of the biggest challenges for any organization is getting control of an explosion of unstructured content, much of it siloed and unanalyzed, and harnessing that information to make smarter decisions and achieve greater efficiency. An information agenda focused on aligning business with IT to achieve both short-term tactical and long-term strategic changes can help drive the needed change. And getting your arms around e-mail challenges is a critical step.

Consistent, reliable and automated categorization of e-mail and other unstructured content is a critical element of the foundation to bringing all e-mail content under management and into compliance quickly. Once the content is accurately cataloged, it can provide a lower risk, lower costs and improved downstream efficiency. Accurately and efficiently categorizing e-mail behind the scenes significantly reduces the risk and burden to end users and makes the discovery of business e-mail more effective, responsive and economic.

Highlights

An automated content system is more likely to be accurate, searchable, accessible, reusable and easier to manage over the information lifecycle.

Not only can knowledge workers be more productive because they're wasting less time looking for information, but authoring experts can be freed from manually tagging and categorizing content. An automated content system is more likely to be accurate, searchable, accessible, reusable and easier to manage over the information lifecycle. It's all part of an agile enterprise content management (ECM) approach that empowers users and improves the organization's ability to react to changing business needs. In short, it's a smarter way to manage content, from e-mail to electronic documents to other content across the agency. Automated techniques increase the consistency of information categorization, which increases reliability. Human beings can be distracted and quickly lose interest in repetitive tasks. And when groups of people get together to organize things, differences of opinion quickly arise, adding to the inconsistent nature of human-performed categorization.

How ICM software helps government agencies define their records policies

According to the new Federal Rules of Civil Procedure (FRCP) and other mandated drivers and policies, if content is electronic and accessible, it is discoverable. In fact, records management needs to include not only e-mail and electronic documents, but all business-relevant information.

ICM software provides consistent, accurate categorization of content, also analyzing it to determine its value to the organization.

IBM Classification Module (ICM) software can provide consistent, accurate and reliable categorization of content, analyzing it to determine its value to the organization. ICM software helps to smoothly incorporate new content with content already under management. It incorporates with and standardizes on the ECM platform from IBM to better manage unstructured content. The software is especially suited to identifying and categorizing content that qualifies as federal records and must be kept for compliance purposes.

Highlights

Later in this paper, we will look at how the U.S. Army, one of the world's largest government organizations, used auto-categorization with ICM software to begin to bring millions of e-mail records under compliance with NARA regulations.

Elements of an auto-categorization system

How do organizations add auto-categorization to the ECM system? Advanced, training-based auto-categorization processes more or less include the basic elements below.

The corpus

The more examples that you have in each category of a corpus, the better your end results will be.

A corpus (plural *corpora*) is a large and structured collection of data, such as e-mail, documents or other texts. The corpus information is typically broken down into categories. Every category in a corpus must contain at least one item, and ideally 20–50 items per category. The more examples that you have in each category, the better your end results can be. A realistic ratio or range of items per category can also provide better results.

Corpora can be created through a variety of means. Frequently, customers have at least some content manually categorized and residing in a central repository. This content can be easily leveraged as a ready-made corpus. Other times, organizations rely upon individuals who are knowledgeable about the information and the records policies of an organization and are asked to create the corpus of documents by manually categorizing content. In building a corpus, consistency is important. There must be agreement on what information gets placed in each category and the reasons why. Failure to provide this consistency can negatively impact the results and provide variations or inconsistencies in the knowledge base, rendering the categorization against new content inaccurate.

Highlights

Once a corpus is created, ICM software can create a knowledge base, statistically analyze unstructured content and categorize that data by context.

Once collected, the corpus serves as a data training set that helps advanced content categorization software automatically determine categorization policies.

A knowledge base

In creating a knowledge base, ICM can take the corpus and begin to create profiles for each category. The software converts the information from a human-readable format into data that can be analyzed by the categorization program. Using the training set established with the corpus, the software can automatically generate a profile of each potential category defined by an organization. It can also prune the content of any extraneous information, using natural language processing to filter out everything except a prioritized list of essential words.

Statistical analysis

Once the knowledge base is established, ICM software uses it as a baseline to analyze uncategorized, unstructured content in an organization. Using statistical analysis, the software compares new documents to the profiles of the categories in the knowledge base. Through this method, it can accurately place the information in the correct category.

Categorization by context

The result of this training-based method is that the ICM software can assign metadata based on the full context of the document. In other words, ICM doesn't just search for a single word or phrase, but analyzes the entire document, discerning the topics in the text. Then, through statistical comparison, ICM categorizes the text by topic. Although the software can execute preconfigured rules, its ability to take the context of the whole document into account provides higher levels of accuracy than more simplistic approaches.

Highlights

Two e-mails can have the same keywords and metadata but different business value — ICM uses context-sensitive analysis to tell the difference.

For example, in figure 1 there are two e-mail messages: one with limited business value and one with clear business value. If we're trying to make a decision on which e-mail to archive and how to file it, a rules-based approach could have difficulty making these distinctions. The two e-mail messages share many metadata values, like sender and recipient. They share some common keywords that might be used to define rules about how to file these messages.

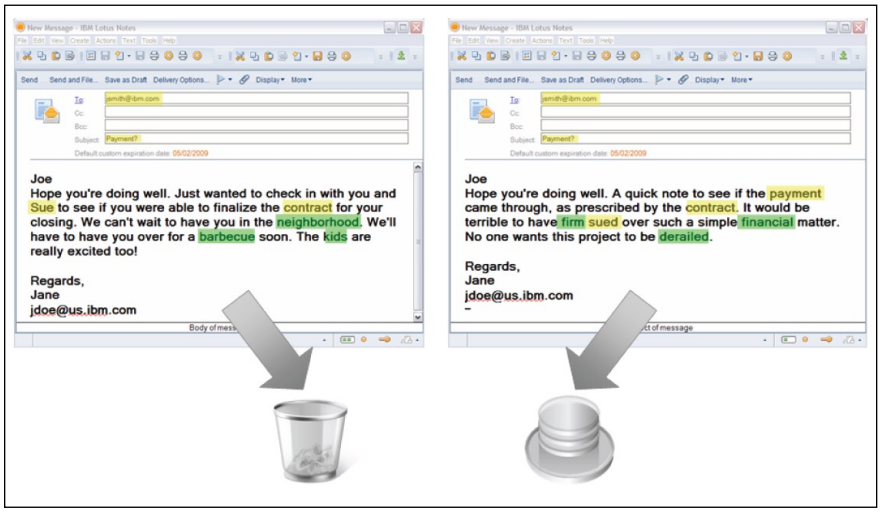

Figure 1: ICM software uses context-sensitive analysis to tell the difference between e-mail with business value and without.

On the other hand, a context-sensitive approach can take into account the full context of these e-mail messages, and factor in the full language being used by the two messages — not just a few keywords or metadata. Therefore, with context-sensitive analysis, ICM software can determine that the e-mail on the left is an inconsequential note between two friends about a barbecue — and that the e-mail on the right concerns an impending legal matter and needs to be retained and saved as a corporate record. It also accounts for misspellings, abbreviations, shorthand and technical terms.

Highlights

Of course rules-based approaches to classifying content will frequently have a role, even when advanced methods are being used. The ICM software provides the capability to combine advanced training-based methods of classifying with rules-based classification. IBM advocates for the combination of multiple methods of categorization analysis in order to realize optimal automated, classification accuracy. For this purpose, the ICM software can also provide the ability to classify documents and e-mail based on the existence of keywords, on the proximity of two words to each other or even on the existence of defined patterns, like social security and phone numbers.

ICM software "learns" from feedback, adapting in realtime to any user changes to content categorizations.

Intelligent software that "learns" as it goes

ICM software also has the ability to "learn" from any feedback provided to it by users. It's a smarter, more responsive way to organize information. As users override or suggest changes to content categorizations through their normal course of business, the system adapts its training and understanding in realtime. So during the very next categorization request, the solution can use what it has learned from previous actions. More recent teachings have more relevance than older ones, allowing the system to adapt and evolve its understanding of how to categorize content as the business adapts and evolves.

The solution can also review content items that don't fit into the current records file plan or category structure and in turn provide category suggestions in a clear, prioritized list. As administrators gain confidence in the quality of decisions being made by the solution, manual auditing becomes less necessary.

Let's dive down a little deeper into how IBM categorization technology works. This service is a weighted combination of analytics, linguistics and statistical methods and tools.

Highlights

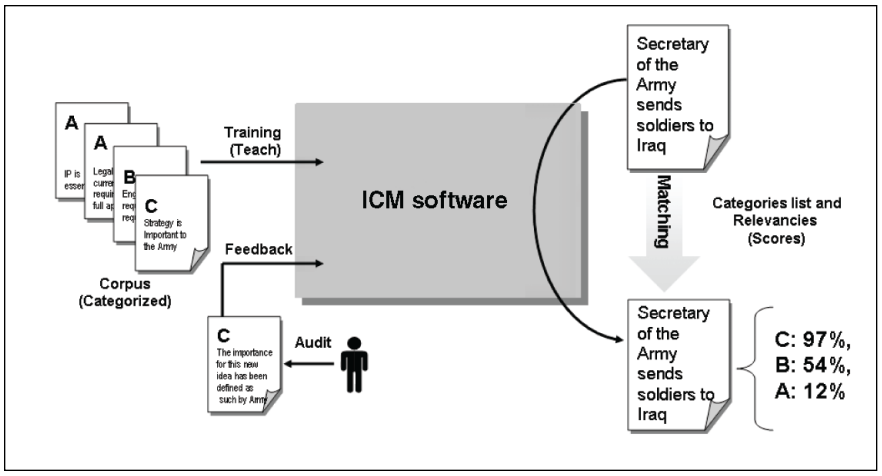

Figure 2: An overview of how ICM software analyzes content for categorization, and how it "learns" and evolves.

Extra rules-based analysis can also be run at any point during context-sensitive categorization.

On the right-hand side of figure 2, a document (which could be an e-mail, a desktop document or any kind of file with long-form text or metadata) is sent to the categorization module and the text and existing metadata are "read." The software runs the content through natural language processing to identify the full set of concepts involved in the document, and a concept profile is compared against the training set for statistical similarity. The content is then assigned a category or categories that are most similar. Extra rules-based analysis can be run at any point. In turn, categories (or a set of suggested categories) are returned for use by the software.

Each category is paired with a confidence level — the higher confidence level you require of automated action, the lower the amount of automation.

Further, each category recommendation is paired with a confidence level. This confidence level is used in a variety of ways. Primarily, it is used to set a level of automation. The higher confidence level you require of automated action, the lower the amount of automation will be. This confidence level can be used to regulate or determine the levels of automation.

Highlights

Moving to the left-hand side of figure 2, we see that the system itself is trained using real content, which is associated with a category in the taxonomy. Your actual business content is used to create this statistical profile of your taxonomy, and it learns from the best possible examples—your real content. It encompasses the "messiness" of real content and the subtleties that a more rules-based approach might miss. It would take a tremendous amount of resources (both time and staffing) to create and maintain such a set of rules. With actual content, the ICM software gets not only the main topic of the document, but also the full context to help differentiate similar categories.

ICM software understands not only the words used in unstructured content, but also the context of the language, as well as associated metadata.

The capability of understanding the meaning of unstructured text and adapting to changing environments in realtime is what makes ICM software noteworthy. The technology understands not only the words used, but also the context of the language, as well as associated metadata. Unlike other technologies, ICM software learns from user interactions and becomes more accurate over time, without requiring explicit document-by-document involvement.

ICM technology currently supports language processing in 16 languages: English, French, Spanish, Italian, German, Portuguese, Dutch, Swedish, Russian, Japanese, Chinese (traditional and simple), Korean, Arabic, Farsi and Hebrew. It can also execute categorization for content written in other languages using a generic language option.

Highlights

ICM software in action—the U.S. Army

Consider the example of the U.S. Army. No one was more familiar with the challenges of records management and compliance to the nine NARA requirements, as outlined in the GAO report, than the Army. With millions of e-mail messages going through the system every year, keeping track of each one and deciding which ones to retain was a challenge. However, the Army had also seen other federal agencies held responsible for not having appropriate records-keeping practices, or at least not adhering to a well-defined records policy. One agency in particular is spending US$5 million a year in discovery of e-mail. What's more, that agency is also under court order to not delete any archives, backups or other records as a result of not having or enforcing a records policy.

To ensure compliance with NARA requirements, the Army needed an effective, efficient records policy that addressed special challenges, such as the participation of active-duty soldiers in combat situations.

The Army saw that it needed an effective, efficient records policy of its own that would help it avoid these pitfalls. But the Army had its own special challenges for e-mail management. Relying on active soldiers to understand records management and file their e-mail or documents appropriately is not practical, especially when the soldiers' primary duties are to support combat operations. Therefore, the Army concluded very quickly that a system needed to be put into place to achieve compliance without user interaction.

The challenges and needs of the Army went beyond simple e-mail management; they included all of their ECM projects.

More generally, the Army's enterprise information architects saw that these types of problems spanned all of their ECM projects, not just e-mail management. The pilot could also help them realize compliance at an enterprise level. To achieve this, their ECM strategy needed to include standardization of content under a single set of rules and policies. But standardization raises a unique challenge—namely, how to manage content created under widely different metadata structures (such as different taxonomies) that are spread across multiple departments, geographies, repositories and applications. These considerations contributed to the Army's desire to have this pilot categorize e-mail across a sampling of offices throughout the Army enterprise.

Highlights

ICM software fit into the Army's current ECM architecture with no development required.

For example, ICM could also solve other categorization challenges for the Army, including:

- *At least 22TB of new content that needed to be accurately cataloged for integration into the ECM platform.*
- *Information already brought under management that was not categorized. This was because either the capability to quickly categorize didn't exist at ingestion time, or the organization didn't have the proper categorization defined or tools deployed. The Army needed the capability to recategorize this existing content to existing (backlog or archive) or newly defined (go-forward) taxonomies.*
- *Agencies that needed reorganization or a Base Realignment and Closure (BRAC), but had two or more conflicting taxonomies that needed to be rectified.*
- *Taxonomies distributed throughout the Army that didn't conform to a standardized structure. The Army needed to normalize its standard taxonomy and make the taxonomy correlate to the official Army file plan.*
- *E-mail that either must be managed as an official record, or that is transitory. The Army needed to institute a records management system that leverages auditable policies and can quickly and consistently assign e-mail records to the appropriate record category and file plan.*

Highlights

The Army and IBM create an ICM pilot program

When the Army approached IBM to help put together a solution to solve this problem, IBM was eager to lend its leadership in ECM compliance solutions. In fact, IBM was able to help by using existing software and hardware assets, and by adding one additional component—ICM software. What's more, ICM fit into the Army's current architecture with no development required. By adding ICM software to its ECM platform, the Army hoped to address its GAO policy and guideline requirements.

A pilot, based on ICM software, was set up to test the efficacy and results of auto-categorization on small sample groups. The Army has a well-defined set of approximately 3,000 records categories. To narrow the scope of this pilot, the Army selected 16 offices for rollout. These offices were chosen to provide a comprehensive sampling of records across the Army. These 16 offices were then cross-referenced with the 3,000 records categories, and the Army determined that 314 records categories should apply to these selected offices. The pilot focused on approximately 400 e-mail users, with plans to roll it out across the Army if successful.

Conducted in three phases, a pilot program was rolled out to test auto-categorization on small sample groups of e-mail.

This pilot program for the new system was conducted in three identical phases. Each phase included e-mail samples from randomly identified Army personnel and was conducted at 30-day intervals, so the team could gather a new e-mail sampling for each test. In the prehase activity, additional effort was required to build the corpus. Records Management and Declassification Agency (RMDA) records manager personnel gathered the e-mail samples, which were used to train the auto-categorization engine. After this first step was completed, Phases I, II and III consisted of running the auto-categorization engine against the additional 30-day samples of e-mail.

Highlights

System reporting and audits were run to determine the success rate of the system's category assignments to the e-mail. The records managers were also involved in reviewing and validating the auto-categorization results throughout the audit process. The objective of this pilot was for 90 percent of e-mail and associated attachments to be correctly categorized by the end of Phase III.

To help enhance the security and privacy of all pilot participants, Nonsecure Internet Protocol Router Network (NIPRNet) and virtual private network (VPN) technologies were used during all three phases to help protect the e-mail samples. The system was accessed only by RMDA records managers once Army e-mail samples were imported into the pilot system. Instead of placing the sampled e-mail on live Army e-mail servers, the messages were handled in .pst files (Microsoft® Exchange e-mail server mailbox archives), helping to provide further protection.

Taking the steps toward auto-categorization

In conducting the ICM pilot program, the U.S. Army took the following steps to achieve auto-categorization in its e-mail samples.

Step 1—The file plan

First, the Army supplied a file plan for selected records in XML format, which was then imported into IBM FileNet Records Manager software.

The Army supplied the file plan for these 314 records categories in XML format, and IBM was able to easily import the XML file plan into IBM FileNet® Records Manager software. Once the records categories were identified, the e-mail could be organized into a format that the ICM software could use to build the corpus.

Highlights

Next, four Army records managers worked with two IBM technical specialists to organize a corpus.

Step 2—The corpus

The corpus had to be built by individuals who understand the Army's file plan and records categories. The Army provided four Army records managers, skilled and well practiced in information categorization and management, who teamed with two IBM technical specialists to organize the corpus. The Army records managers provided the subject matter expertise for the Army file plan as well as the business knowledge required. The IBM specialists provided expertise in the software and in the management of the system, and they helped facilitate discussions between the records managers.

Because the Army used a Microsoft Exchange e-mail server, and the records managers were already familiar with Microsoft Outlook software, IBM configured the Microsoft Outlook software to review the sample e-mail messages. The specialists also configured Microsoft Outlook software with folders representing the categories they expected to see most often from each of the Army offices participating in the pilot.

The individuals collecting data agreed on definitions to ensure consistency in categorization.

The records managers collaborated on categorizing the e-mail. The IBM staff observed their interaction and information-gathering techniques. Each records manager was responsible for identifying e-mail from his or her assigned offices. At the same time, to realize the benefits of consistent analysis with ICM software, the records managers needed to agree on definitions. The individuals collecting the information and organizing it into the corpus must use consistent approaches. There must be agreement on what information gets placed in each category and the reasons why. Failure to provide this consistency could negatively affect the results and provide variations or inconsistencies in the knowledge base, and thus render the categorization against new content inaccurate.

Highlights

The records managers had to read the e-mail and drag each message into the appropriate record category. Microsoft Outlook software shows a subset of Army records categories configured as folders. IBM and the Army used this simple organization of e-mail into folders as the defined corpus. E-mail without business value, such as external marketing and informal correspondence, went into a folder marked "Non Records."

Collaboration between records managers and IBM staff helped IBM understand the corpus results and how to best adjust for accuracy.

The observation of this process was probably the most important aspect of the pilot. For the first three days, the records managers collaborated to discuss the e-mail to decide how each message should be categorized. This simple collaboration allowed the records managers to better understand one another's opinions. More importantly, it helped IBM understand the results that were produced in building the corpus and how to make adjustments so that the final results could also be adjusted to provide better accuracy.

Building the corpus took approximately three weeks. At the end of the building process, the Army had a corpus that contained the following:

- *11,915 e-mail messages*
- *54 records categories (folders) identified as being associated with the assigned offices*
- *28 categories with 15 or more examples*
- *14 of those categories with more than 100 examples*
- *4 of those categories with more than 1,000 examples*
- *26 categories with 14 or fewer examples*

Highlights

Using tools within ICM software, the team was able to take raw information and adjust for inconsistencies that arose from having multiple people categorize e-mails.

Step 3—The knowledge base

With the reporting tools provided by the ICM software, IBM and the Army were able see the raw information in the corpus and make adjustments for inconsistencies that may have been placed into it by having multiple people work on it. There were inevitably a few inconsistencies. However, they were natural, auditable, recognized and corrected. The ICM module had several tools that enabled the team to gain better insight into information contained within the corpus.

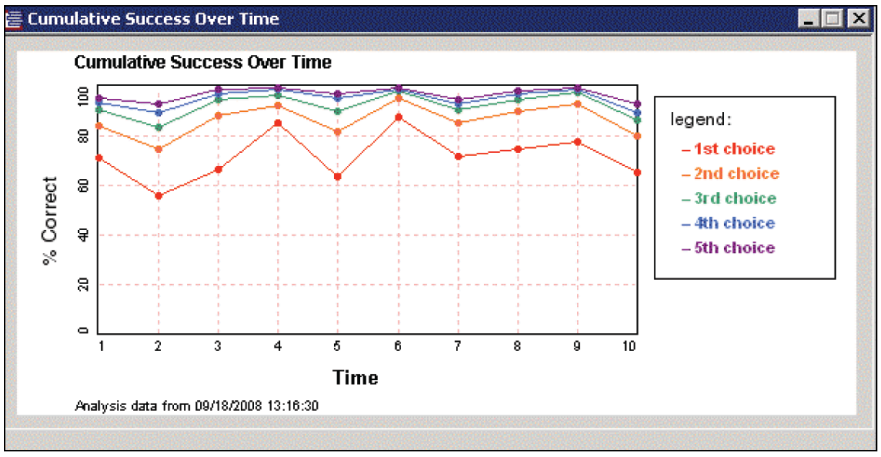

Figure 3: With the tools in the ICM software, the team was able to track the categorization engine's success over time and account for inconsistencies.

One tool enables users to look at the cumulative success of the categorization tool over time. In figure 3, the raw information shows that the categorization was not consistent with its success as time progressed. By looking at this information and discussing it, the team was able to isolate two primary reasons for the inconsistency—different interpretations of the Army file plan by different records managers, and the granularity of the Army records categories. The

Highlights

A tool that tracked cumulative success over time showed that accuracy improved when some e-mails were placed in two categories instead of just one.

Army had been dealing with these types of records categorization inconsistency issues in the past with manual categorization techniques. In meetings with Army management, the team discussed these issues and determined that in the past, the Army resolved inconsistencies by categorizing the record in two or more categories. Having already anticipated these issues, the Army wanted to use the ICM software to categorize items in two or more records categories. This helps alleviate the inconsistencies, or in this case, differences in opinions, because both parties had relevant reasons for placing the information in different categories. Upon further investigation from the IBM staff, both categories actually made sense, given the context of the e-mail. With this capability, the cumulative success over time and the total precision versus recall charts improved dramatically.

According to the chart in figure 4, where the items are categorized in two records categories, the effort had a much more consistent success rate over time.

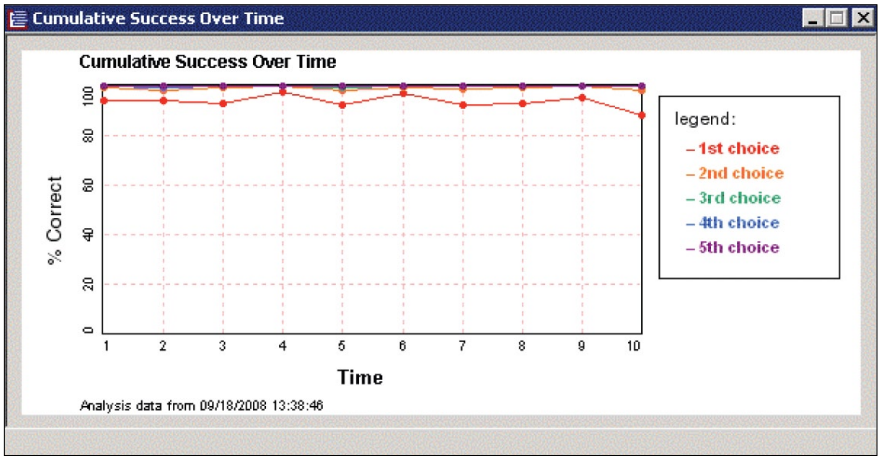

Figure 4: A graph shows the cumulative success of the categorization when items are categorized in two records categories.

Highlights

The precision of categorization also improved with adjustments.

We can also see by comparing figures 5 and 6 how the precision of categorization also increased by making these adjustments. Figure 5 shows the precision versus recall before the categorization adjustments were made. Figure 6 shows the results after the adjustments.

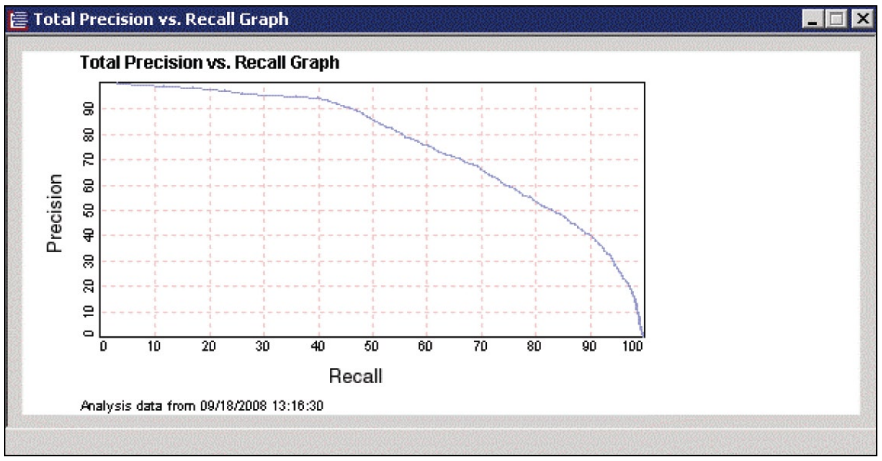

Figure 5: Total precision versus recall results before items were placed into two records categories

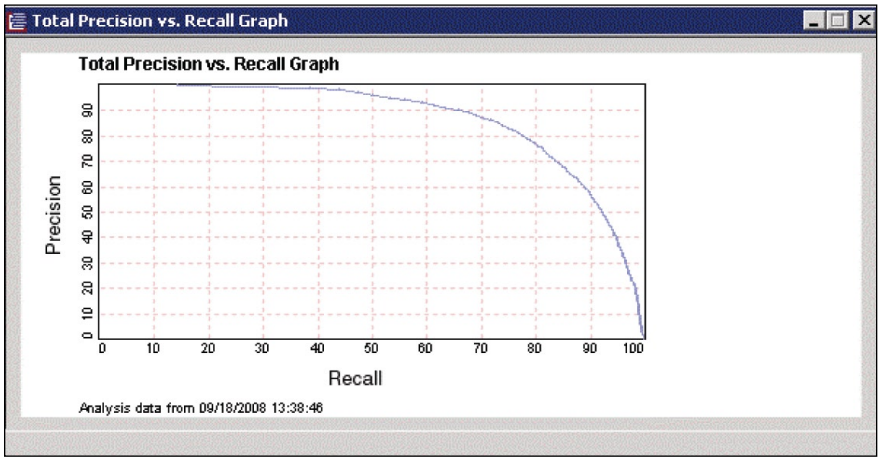

Figure 6: Total precision versus recall results after the items were categorized into two areas

Highlights

ICM software incorporated feedback from records managers in realtime, improving accuracy.

Feedback in realtime

Through the course of reviewing and auditing, the Army records managers provided feedback to the ICM software. The tool then processed that feedback and adapted to it in realtime—the very next categorization request learned from the feedback provided. More recent feedback is weighted higher than older information, allowing the system to adapt and evolve its understanding of how to categorize content.

Results of the Army ICM pilot program

Before the Phase I audit, 84 percent of the e-mail sample was categorized. Phase I results show that after the audit, 98 percent of the e-mail messages were categorized, indicating that the ICM tool was doing a very good job. And in the areas where categorization was not working, the issues were easily identifiable and correctable. With feedback, the ICM software was able to make more accurate assessments, as shown in the postaudit results. The other two phases of the pilot yielded similarly positive results as shown in table 1 below.

Table 1: Comparing categorization results preaudit and postaudit for all pilot phases

	Phase I	Phase II	Phase III
Total number of e-mail messages processed	581,634	581,526	735,333
Preaudit			
Categorized	84.5%	99.1%	98.4%
Assigned to a record category	21.3%	10.6%	29.8%
Assigned as a nonrecord	63.2%	89.4%	70.2%
Not categorized	15.5%	15.9%	1.6%
Postaudit			
Categorized	98.8%	99.99%	99.99%
Assigned to a record category	22.3%	10.9%	30.4%
Assigned as a nonrecord	63.2%	89.1%	69.6%
Not categorized	1.2%	0.01%	0.01%

Highlights

With categorization of e-mail, information became more accessible, discovery became easier, and the Army saved both time and money.

Conclusion

As the results indicate, the pilot objectives were exceeded in each of the three phases. These results were a product of the dedication and determination of the entire pilot team of Army and IBM personnel. The Army records managers, as well as their management, dedicated their time to the success of this pilot, building understanding and making adjustments based on feedback from the ICM engine.

By categorizing these e-mail messages, information became much more accessible and meaningful. Discovery of information became much easier, saving time and costs in a variety of business aspects—including legal discovery, information reuse and information relevancy. With ICM software, the Army realized its goal of more effective categorization and auditing of records with less user interaction.

But the Army also achieved another, equally important, goal—storage cost savings. If we look at the above results, approximately 500,000 e-mail messages per 400 users were nonrecords in each phase. We can calculate that 500,000 e-mail messages at 50,000 bytes of disk space per e-mail amount to 25GB of storage per month. If we divide that figure by 400 users, results indicate that the Army could be saving 62.5MB per user per month in storage space. The Army has approximately 1.2 million users on e-mail; therefore, it can be concluded that over a year, the Army could save 900TB of storage space by eliminating non-business-related e-mail. If we consider that 1TB of storage costs approximately US$2,000, the Army can save US$1.8 million per year in storage costs alone.

Highlights

Storage cost savings are only the beginning. By categorizing these e-mail messages, information becomes much more accessible and meaningful to users—it becomes intelligence. And discovery of information becomes much easier, with information on demand saving time and costs in a variety of business aspects such as legal discovery, information reuse and relevant information discovery.

"As a records manager with a 25-year background in federal and civilian records management, I believe the automatic categorization of information is the next logical evolution in managing the records of an organization."

—Brenda Fletcher, records manager, United States Army

IBM provides ECM solutions for more than 13,000 global companies.

Why IBM?

The IBM enterprise content management operation helps the world's top companies make better decisions, faster. As a marketplace leader in content, process and compliance software, IBM can deliver a broad set of mission-critical ECM solutions that help solve today's most difficult business challenges: managing unstructured content, optimizing business processes and helping to satisfy complex compliance requirements through an integrated information infrastructure. ECM solutions also provide an entry point for realizing the information on demand vision. More than 13,000 global companies, organizations and governments rely on ECM solutions from IBM to improve performance and remain competitive through innovation.

As you work to gain greater business value from the information assets spread across your enterprise content repositories, just figuring out what you have can be a struggle. ICM software is part of a portfolio of security-rich and scalable enterprise search and discovery solutions—including solutions for automated classification, unstructured document search and content analysis—that can help you examine and classify information assets company wide. Ask your IBM representative how other solutions in the search and discovery portfolio from IBM can complement your environment to support better, faster insights and business decisions.

For more information

To learn more about ICM software and other enterprise content management solutions from IBM, contact your IBM representative or IBM Business Partner, or visit:

ibm.com/software/data/content-management

To find out more about compliance and discovery, visit:

compliancewarehouse.techweb.com

To find out more about GAO report GAO-08-742, visit:

gao.gov/products/GAO-08-742

Additional reading

"Military Personnel: Army Needs to Better Enforce Requirements and Improve Record Keeping for Soldiers Whose Medical Conditions May Call for Significant Duty Limitations," GAO-08-546 June 10, 2008, *U.S. Government Accountability Office*, http://www.gao.gov/products/GAO-08-546.

Appendix

Transitory records: Under the authority of General Record Schedule 23, Item 7, or a NARA-approved agency records schedule, transitory records have very short-term (180 days or less), NARA-approved retention periods. Agencies may elect to manage such records on the e-mail system itself, without the need to copy the record to a record-keeping system, provided that (1) users do not delete the messages before the expiration of the NARA-approved retention period, and (2) the system's automatic deletion rules ensure preservation of the records until the expiration of the NARA-approved retention period.

File plan: A file plan specifies how records are organized hierarchically in a records management environment. A file plan is similar to a collection of containers; a container represents a holding place into which you place records related to a common subject or theme, or another container, together. File plans are also used for defining records security and retention rules (from the schedule) against containers.

Taxonomy: Taxonomy is the practice and science of classification, or in Army terms, categorization. The word comes from the Greek taxis ("order" or ""arrangement") and nomos ("law" or "science"). Taxonomies, or taxonomic schemes, are composed of taxonomic units known as taxa (singular *taxon*), or kinds of things that are arranged frequently in a hierarchical structure as categories.

IBM Corporation
Software Group
3565 Harbor Boulevard
Costa Mesa, CA 92626-1420
U.S.A.

Produced in the United States of America
April 2009

IMW14166-USEN-00