Classify your enterprise content and manage taxonomies using IBM Classification Module

Standardizing content and associated metadata across the enterprise

Marc Grosse

December 20, 2007

Large organizations need a standard and consistent way to understand and access the unstructured information that resides in their various file systems and content management repositories. IBM® has created a solution to meet this business challenge. IBM Classification Module (ICM) drives business value by removing the need for business and IT users to determine where a piece of content will reside within a managed content repository such as FileNet P8. Even more important, this content can be searched and found quicker resulting in faster business decisions that drive customer satisfaction and value. This article shows you how to classify your enterprise content and manage taxonomies using IBM Classification Module.

Introduction

Within the last few years, many companies have implemented Enterprise Content Management (ECM) systems to manage and deal with the proliferation of unstructured content being created on a daily basis. However, there is still a tremendous amount of content that is not being managed, such as documents on departmental file systems, customer emails and other content repositories. These same companies are trying to develop strategies for creating reliable and consistent access to this content, including developing taxonomies or methodologies for categorizing the content.

Many companies have implemented some form of content classification, but they are often manual and are left to individual departments or business units that have different rules and criteria for the content they are classifying. Often, these classification practices are not succeeding. Finding a way to successfully catalog and classify this content can greatly benefit users who need this information for important decision-making on a daily basis.

How can this be achieved? This can by achieved through the use of the IBM Classification Module. This article introduces the IBM classification and taxonomy management solution in the form of
the IBM Classification Module. This module helps solve the problem by standardizing content and metadata throughout the enterprise.

Defining content classification and taxonomy management

Prior to addressing the technical capabilities of the IBM Classification Module, it is important to understand the key terms for building a solution of this nature.

Today, content stored in a content repository contains unique metadata that describes attributes of the content. A taxonomy is a collection of different metadata values for a particular attribute or category. Examples of taxonomies include the Dewey Decimal System and folders or document classes within FileNet P8 or other content management systems.

The art of determining where a piece of content should reside in a taxonomy is called classification. It provides an organization with a manual or automated way of determining the context of content. Within ECM, taxonomies ensure that content is accurately catalogued and easily accessible; having accessible taxonomies is one of the many business benefits of ECM. When content is presented in a formal, managed environment, users profit from having easier access and are able to make quicker, more informed decisions.

Critical business drivers

Taxonomy and classification management is a common problem that many enterprise architects face today. They witness the proliferation of content in their organization, and their goal is to standardize the content under a single set of rules or policies. There are several major areas where an automated classification system can be applied. These include:

- Generic content classification - Standardizing on an ECM repository such as FileNet P8 can be an arduous task as it requires the ability to bulk load content and bring it under new management. This content needs to be accurately cataloged.
- Existing content under ECM - Existing content under management may not be catalogued or classified accurately. Often, it has been done manually by users and varies widely within different business units. This content has to be reclassified to adhere to a standard informational model.
- Merger and acquisitions - Two companies who have been recently merged together have two separate and distinct content management repositories and taxonomies. These taxonomies need to be normalized, and the respective content has to be reclassified to address a standard taxonomy.
- Automated decision making in a business process - Classification can be used to determine the proper routing of a workflow as well as enhance the efficiency of a workflow by populating attributes to allow knowledge workers to improve the speed and efficiency of their decision making.
- Email automation in a call center environment -- Classification can be effective at increasing deflection rates for handling email by automatically sending out canned responses to customer inquiries.
- Business user analytics that provides insight into search-based merchandising and product catalog offerings
• Support for WebSphere Commerce e-Sites and contracts

Automated classification is seen as a cost-effective way to ease the ECM standardization process ensuring that the content within enterprise repositories are accurately catalogued, and more importantly, findable and usable by enterprise users.

**Technology overview**

The foundation of IBM taxonomy management solution is the IBM ECM Classification. This product leverages natural language processing (NLP) and semantic analysis for interpreting different bodies of text and applying context about this text within a taxonomy. It combines text-analytics and rules that read and understand ingested content and assigns metadata based on the engine’s perceived understanding of the content. The output of this assigned logic can automate appropriate, previously human-centric actions such as categorizing, filtering, routing, notifying, and responding on documents, emails and customer interactions.

The capability of understanding the meaning of unstructured text and adapting to hanging environments in real-time is what makes ICM unique. The technology understands not only the words used, but also the context of the language, as well as associated metadata. Unlike other technologies, ICM self-learns, becoming more accurate over time with limited human intervention. This self-learning allows the engine to adapt and evolve in its understanding of how to classify content as the business adapts and evolves. From a taxonomy management perspective, the solution also reviews existing catalogued content and suggests new groups and categories to augment the existing taxonomy or create an entirely new taxonomy.

**How classification works**

The workflow diagram below illustrates how the ICM classifies content. On the right hand side of the diagram, a document is consumed by the engine, the text is analyzed and a set of suggested categories are returned for use by any ECM System. The ICM Relationship Modeling Engine (RME) uses a two-phase process to classify input texts. First, the NLP engine extracts concepts from free-text fields, and generates a Semantic Modeling Language Document (SML). Concepts are information that can give ICM RME a hint about the text. Then, the Semantic Modeling engine performs statistical pattern matching on the SML by comparing it with the content of categories residing in the KB to generate the relevancy scores. Suggested categories and a relevance score are returned. This score represents a confidence factor derived from a concept model that is created at runtime and compared to an existing training set for statistical similarity. The confidence factor, expressed as a percentage, can drive applications to automate various actions such as routing, notifying or responding.

On the left-hand side of the diagram, the system is continuously being trained by content that is being analyzed and categorized into a taxonomy. As the categorized corpus begins to grow, the taxonomy continues to get refined and new categories are created or modified to reflect the existing content corpus. This approach has far better success at categorizing content that typical rule-based engines. More importantly, this process helps to define not only the main topic of the document, but also the full context of the content which helps differentiate similar categories.
How the ICM works with FileNet P8

ICM has been highly integrated with FileNet P8 to provide a highly scalable and robust taxonomy automation solution. The integration architecture consists of four main components:

- **The classifier** executes an initial bulk classification of content at ingestion via FileNet Capture or a reclassification of existing content in P8. Content is either automatically classified into a respective document class or folder, or it is kicked out for manual intervention. A configurable parameter (the auto classify review percentage) determines what percentage of the content being classified will be siphoned off for auditing purposes. Initially, it is expected that this percentage should be configured low to allow for the limited number of content existing in the knowledge base. As more content is classified, the system becomes more accurate in its classification, and, thus, this parameter can be adjusted higher.

- **Existing content under ECM** Existing content under management may not be catalogued or classified accurately. Often, it has been done manually by users and varies widely within different business units. This content has to be reclassified to adhere to a standard informational model.

- **The Classification Review User Interface** is a web-based tool resembling P8 Workplace that allows business analysts or IT users to review content that is not automatically classified. This content is passed to a queue where a reviewer manually classifies the content or audits the classification derived from the engine. In addition, this interface allows for content not stored in a content repository to be ingested and manually classified.

- **The content extractor** is an intuitive tool for extracting both content and associated metadata, document classes and folders from P8 to train the system and enable content classification. This tool serves two main purposes:
  - Builds an existing corpus of content for the classification engine to classify.
  - Provides feedback over time for taxonomy and category refinement.

- **The taxonomy proposer** recommends new categories, or even new taxonomies, based on the sample content extracted by the Content Extractor Tool.
Summary

Large organizations need a standard and consistent way to understand and access the unstructured information that resides in their various file systems and content management repositories. IBM has created a solution to meet this business challenge. ICM drives business value by removing the need for business and IT users to determine where a piece of content will reside within a managed content repository such as FileNet P8. Even more important, this content can be searched and found quicker resulting in faster business decisions that drive customer satisfaction and value.
Related topics

- IBM Classification Module for OmniFind Discovery Edition web site, get the links to the latest ICM documentation (product manuals, white papers, and technotes), Fixpaks, and many other resources.
- FileNet P8 Platform web site: Learn more about the FileNet P8 platform.
- Visit the developerWorks Enterprise Content Management page to read articles and tutorials and connect to other resources to expand your ECM skills.

© Copyright IBM Corporation 2007
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)