GALAXYCONSULTING
Information Management Solutions

Automatic Classification

White Paper
By Galaxy Consulting

April

A

At Your Service…Today…Tomorrow
We Appreciate The Privilege Of Serving You!

January 2015

Abstract

Recent research has shown that two-thirds of organizations cannot access their information assets or find vital enterprise documents because of poor information classification or tagging. The survey suggests that much of the problem may be due to manual tagging of documents with metadata, which can be inconsistent and riddled with errors, if it has been done at all.

Organizations that overcome manual sorting and classification in favor of an automated document classification & processing system can realize a significant reduction in manual entry costs, and improve the speed and turnaround time for document processing.

Benefits of Automatic Classification

An ability to find content in a content management system (CMS) is crucial. One of main goals of having a CMS is to make content easy to find, so you can take an action, make a business decision, do research and development work, etc.

The main challenge to findability is anticipating how users might look for information. That's where categorization comes into play. The quality of the categorization of each piece of content makes or breaks its findability. Theoretically, good tagging will last the lifetime of the content. You would think that if you do it well initially, then you can forget about it until it is time to retire that content. But reality can be very different.

Many issues complicate content categorization. They include:

- sheer volume, velocity, and variety of internal and external-facing content which needs management;
- evolving/emerging regulations and compliance issues, some of which need to be retroactively applied;
- need to limit the company's exposure and to support the strength of its position in any legal activity.

Some organizations face the added challenge of integrating content from acquisitions or mergers, which most likely use content management structure, categorization, and methodologies that are incompatible and of inconsistent quality.

Considering these issues, the success factor for good content categorization is automatic categorization.

Traditionally, keywords, dictionaries, and thesauri are used to categorize content. This type of categorization model poses several problems:

- taxonomy quality - it depends on the initial vision and attention to detail, and whether it has been kept current;
- term creep - initial categorization will not always accommodate where and how the content will be used over time, or predict relevancy beyond its original focus;
- policy evolution - it can't easily apply new or evolving policies, regulations, compliance requirements;
- cost and complexity - it is difficult and costly, if not practically impossible, to retroactively expand the original categorization of the existing content if big amount of content is added.

Automatic Classification

Using technology to automatically categorize content is a solution. It applies the rules more consistently than people do. It does it faster. It frees people from having to do the task, and therefore has fewer costs. And, it can actively or retroactively categorize batches or whole collections of documents.

You can experience these benefits by using concept-based categorization driven by an analytics engine integrated into the content management system. These systems mathematically analyze example documents you provide to calculate concepts that can be used to categorize other documents. Identifying hundreds of keywords per term, they are able to distinguish relevance that escapes keyword and other traditional taxonomy approaches. They are even highly likely to make connections that a person would miss.

Specific benefits of automatic content classification are:

More consistency. It produces the same unbiased results over and over. Might not always be 100% accurate or relevant, but if something goes wrong, it is at least is easy to understand why.

Larger context. Enforces classification from the whole organizations perspective, not the individuals. For example, a person interested in sports might tag an article which mentions a specific player, but forget/not consider a team and a country topic.

Persistent. A person can only handle a certain number of incoming documents per day, whilst an automatic classification works round the clock.

Cost effective. Possible to handle thousands of documents much faster than a person.

Approaches to Classification

Manual - requires individuals to assign each document to one or more categories. It can achieve a high degree of accuracy. However, it is labor intensive and therefore is more costly than automatic classification in the long run.

Rule-based - keywords or Boolean expressions are used to categorize a document. This is typically used when a few words can adequately describe a category. For example, if a collection of medical papers is to be classified according to a disease together with its scientific, common, and alternative names can be used to define the keywords for each category.

Supervised Learning - most approaches to automatic classification require a human expert to initiate a learning process by manually classifying or assigning a number of "training documents" to each category. This classification system first analyzes the statistical occurrences of each concept in the example documents and then constructs a model or "classifier for each category that is used to classify subsequent documents automatically. The system refines its model, in a sense "learning" the categories as documents are processed.

Unsupervised Learning - these systems identify both groups or clusters of related documents as well as the relationship between these clusters. Commonly referred as clustering, this approach eliminates the need for training sets because it does not require a preexisting taxonomy or category structure. However, clustering algorithms are not always good at selecting categories that are intuitive to users. On the other hand, clustering will often expose useful relationships and themes implicit in the collection that might be missed by a manual process. For this reasons, clustering generally works hand-in-hand with supervised learning techniques.

Each of approaches is optimal for a different situation. As a result, classification vendors are moving to support multiple methods.

Most real world implementations combine search, classification, and other techniques such as identifying similar documents to provide a complete information retrieval solution. Organizations having document repositories will generally benefit from a customized taxonomy.

Once documents are clustered, an administrator can first rearrange, expand or collapse the auto-suggested clusters or categories, and then give them intuitive names. The documents in the cluster serve as initial training sets for supervised-learning algorithms that will be used subsequently to refine the categories. The end result is taxonomy and a set of topic models which are fully customized for an organization's needs.

Automatic Classification Tools

There are few taxonomy and automatic classification tools types:

- Thesaurus/ontology management software.
- Other software with thesaurus/taxonomy modules.
- Auto-categorization/text mining software.
- Other software supporting creating taxonomies: mind-mapping or concept modeling tools; card-sorting tools; web analytics.

Thesaurus/Ontology Management Software

Features of this software are:

- Maintains terms and their relationships (equivalencies, hierarchical, and associative): as reciprocals; when renaming, merging, subsuming, or deleting terms;
- disallows invalid relationships (according to standards);
- supports term notes and other attributes for terms;
- supports candidate/approved terms; includes term creation and update dates;
- generates reports in various thesaurus display formats (hierarchical, alphabetical);
- exports data in interoperable formats for importing into a content management, indexing, search, retrieval system;
- supports thesaurus standards: ANSI/NISO Z39.19 or ISO 2788;
- interface design and ease of use;
- multiple taxonomy display options;
- term searching;
- spell-checking;
- speed (limited mouse clicks) for repeated term and relationship additions;
- single-step new term & relationship creation;
- single-step branch (term and narrower terms) moving;
- drag & drop relationship adding;
- user-defined (customizable) relationships;
- user-defined term notes and term attributes;

- bilingual or multilingual taxonomy support;
- importing and exporting formats;
- connectors to enterprise search systems.

Ontology software has these additional features:

- classes for terms;
- customizable semantic relationships between terms (hierarchical and associative), dependent on class
- WC3 or RDF standard outputs (for OWL).

These are few examples of thesaurus management software:

- Data Harmony - Multi-platform java-based (used on Windows, Mac, Solaris, Linux). Client software allows remote access. All standard thesaurus display types can be used as view options. User defined associative and equivalence, but no user-defined hierarchical relationships. Sold separately or combined with M.A.I. (Machine Aided Indexer) as MAIstro. Other software extensions available.

- Synaptica - Web browser-based, priced per user, per year, per vocabulary. 12 graduations of permission levels. Can assign relationship weights. Global term and relationships editor, creating a list of terms to edit. Side-by-side editor with drag-and-drop. Imports: CSV, text, MS Excel, XML (including schemas of ZThes, RDF, SKOS, and OWL). Exports: CSV, HTML, MS Word, MS Excel, XML (including schemas of ZThes, RDF, SKOS, and OWL).

- Semaphore Ontology Manager - Supports creating thesauri according to ISO 2788 standard. Supports creating ontologies, through customizable relationships and user-created classes. User-defined term attributes and metadata. Multiple user access/privilege levels. Imports/export in CSV, XML, Zthes, SQL databases, and MultiTes files. Related products: Classification Server for automated classification. Ontology Service for a navigation system.

- Wordmap - Multi-platform java-based. One of a suite of products including Wordmap Intelligent Text Classifier, Taxonomy Connectors for SharePoint and Endeca. User-defined relationships; can also turn on or off relationship name display. Can display two taxonomies side by side and drag and drop. User access/privileges can be set at the individual node level. Imports: CSV, Excel, XML; Exports: XML. Real-time access: Java API.

- SchemaLogic- Enterprise Suite - Provides thesaurus management according to ANSI/NISO standards, plus broader structural metadata support. Can create customizable relationships. Can assign various permission levels to vocabularies or terms. Classification module supports 3rd party auto-indexing. Connectors to SharePoint, EMC Documentum, and FAST ESP. Can import CSV or XML files.

- STAR/Thesaurus - Stand-alone or integrates with STAR family of products for records management, collections management, archives management, DAM. Supports standard thesaurus relationship but not customizable relationships. Supports unlimited user-defined notes and categories. Various output report display formats. Import/export ASCII text and CSV, but not XML.

- SoutronTHESAURUS - Markets in the U.S. through partnership with InMagic. Stand-alone or integrates with SoutronGLOBAL or SoutronSOLO library management systems, or with InMagic Presto social knowledge management software. Supports standard thesaurus and user-defined relationships. Supports term merging. Imports from XML; exports to XML or CSV.

- Mondeca - ITM T3 (Intelligent Topic Manager: Thesaurus, Taxonomies, Terminologies) - Since 2008, addition to Intelligent Topic Manager set of products for knowledge management, semantic portals, and e-catalogs. Web-based collaborative application. Conforms to both SKOS vocabularies and OWL-

standard ontologies. Connectors to text mining, classification, and search tools. Imports/exports XML, RDF, and SKOS.

These are few examples of ontology software: TopBraid Composer, Altova SemanticWorks, Protégé, SMORE, SWOOP, CMAP Tools Ontology Editor.

Auto-categorization/Text Mining Software

Features of auto-categorization software are:

- Algorithms, statistics, and training documents – utilize a large set a sample documents per taxonomy term to "train" the system to learn to index.
- Rules based – generate and edit or write "rules" for each term based on co-existing words, proximity, Boolean logic, etc.

In addition to these features, text mining software extracts relevant terms from texts to generate candidate taxonomy or supplement an existing taxonomy.

Auto-categorization, text mining, and search systems that utilize taxonomies handle these taxonomies in different ways:

- with pre-installed taxonomies that cannot be edited;
- with pre-installed taxonomies that the user may edit and extend through the user interface;
- automatically generate a taxonomy that can be edited;
- support the import of taxonomies but do not support the editing of those taxonomies;
- support the import of taxonomies and then the editing of those taxonomies;
- various combinations of above.

Software that can import and use taxonomies but lacking user-interface features to edit those taxonomies includes: Microsoft SharePoint, IBM Classification Module, Endeca, Temis, Vivisimo, Mindbreeze, Exalead, PerfectSearch.

They collaborate with other vendors that develop taxonomies and/or have taxonomy editing capabilities.

Examples of tools with *some* taxonomy management capabilities: Inxight SmartDiscovery Analysis Server, Autonomy Collaborative Classifier, Autonomy Interwoven MetaTagger, Lexalytics Classifier, Conceptsearching.

Few examples of auto-categorization tools with full thesaurus management capabilities:

- Data Harmony MAIstro (combines Data Harmony Thesaurus Master and Machine-Aided Indexer) - Automatically creates a basic rule for every term and its variants in the Thesaurus Master's thesaurus. Rules may be edited and additional rules can be manually written statistics module tracks the editor's term choices and compares them with M.A.I. term suggestions, sorting them as hits, misses, and noise to guide and prioritize the editor's fine-tuning of rules. Can be used for machine-aided indexing or fully automated indexing. Connectors to Sharepoint and search engines.

- Smartlogic Semaphore Classification Server (connects with Semaphore Ontology Manager) - Creates classification rules directly from a taxonomy/thesaurus/ontology, and applies these rules to content as it is received to automatically classify content. Rules are based on the term, its equivalencies, and broader/ narrower/related terms. Employs 20 different kinds of rules. Rules have weights and scores. Variants based on spelling, plurals, and stemming may also be considered. Manual rules can take precedence over generated rules.

- Wordmap - Intelligent Text Classifier (connects with Wordmap Taxonomy Manager for leveraging thesaurus) - Auto-classification based on statistical method based of Support Vector Machine algorithms

and machine learning with training documents. Pre-packaged with statistical algorithms based on a generic taxonomy, the U.K.'s Integrated Public Services Vocabulary (IPSV), for which each of hundreds of terms have already been "trained" with representative documents. Wordmap also offers Taxonomy Connectors for taxonomy-driven tagging and search within SharePoint and Endeca.

- SAS Enterprise Content Categorization (formerly Teragram TK240) - Supports taxonomy building or connects with SAS Ontology Manager. ECC supports equivalent, hierarchical & related relationships. Ontology Manager supports customized relationships and attributes. Utilizes both auto-categorization and entity/concept extraction. Auto-categorization bases on rules. Rules-writing supported with a graphical tree view of Boolean operators and commands. User can define weighting of terms.

- OpenText Nstein Text Mining Engine (TME) - modules include concept extractor, entity extractor, auto-categorizer, automated abstract creation, sentiment analysis. Taxonomy Manager Module (not sold separately) supports creating & editing hierarchical, associative and equivalence relationships according to ANSI/NISO standard. Auto-categorization technology based on use of training sets for taxonomy terms, combined with concept extraction technology. Ships with pre-installed taxonomies already "trained" for auto-categorization.

- SAS Text Miner provides tools that enable you to extract information from a collection of text documents and uncover the themes and concepts that are concealed in them. In addition, you can combine quantitative variables with unstructured text and thereby incorporate text mining with other traditional data mining techniques. SAS Text Miner is a component of SAS Enterprise Miner. SAS Enterprise Miner must be installed on the same machine.

Some taxonomy tools are stronger in taxonomy/thesaurus/ontology management. Some taxonomy tools are stronger in auto-categorization. A few tools combine both, but vendor partnerships and connectors can also achieve high results.

Conclusion

An effectively managed content delivers better cost of content management and reduced exposure to risk. While this alone is reason to implement improvements in categorization, there are other reasons.

Superior categorization through conceptual analysis also affects operational efficiency by enabling fast, accurate, and complete content gathering. A significant benefit for any enterprise is that it allows more time for actual work by reducing the time it takes to find necessary information. It is of critical importance for companies whose revenue depends on their customers quickly and easily finding quality information.

Conceptual analytics systems deliver two other advantages over traditional taxonomy methods and manual categorization. It creates a mathematical index, so it is useless to anyone trying to discover private information or clues about the company. Also, it is deterministic and repeatable. It will give the same result every time and so it is very valuable in legal or regulatory activities.

Concept-based analysis makes content findable and actionable, regardless of language, by automatically categorizing it based on understanding developed from example documents you provide. Both internally and externally, the company becomes more competitive with one of its most important assets which is unstructured information.

About Galaxy Consulting

Galaxy Consulting provides services in business analysis and usability, content and knowledge management, records management, information architecture, enterprise search, taxonomy development and management, document control, and information governance.

Galaxy Consulting was founded with the mission and vision of helping organizations to manage their valuable information assets. Many of our clients, both large and small, have dramatically improved efficiency and reduced unnecessary labor hours through efficient methods, processes, and solutions we created.

Galaxy Consulting believes in partnerships with our clients. We are committed to working with you and to helping you transform your business. We will increase efficiency and productivity, maintain regulatory and legal compliance, improve collaboration, enhance innovation, and reduce costs through effective information management!

Call us TODAY to schedule a free, no obligation consultation!

Contact Us

Office: 650-474-0955 Info@galaxyconsulting.net

Mobile: 650-716-3609 www.galaxyconsulting.net