

Criteria for Selection: Classification Technologies

Kathleen Hall and Daniel W. Rasmus

Giga Position

A new set of tools and technologies has emerged to aid in developing and maintaining content taxonomies and classifying content. This market segment will become increasingly important as organizations continue to struggle with vast amounts of unstructured content. Simple search is no longer sufficient. Content taxonomies and consistently classified documents can help improve search quality, provide alternative access points to content and enable a better understanding of what content exists within an organization.

Unstructured content must move from random collections to semistructured collections, with structure stored in metadata. In most cases, that metadata will be derived from the content itself via automated classification tools. The ever-expanding amount of unstructured content in many companies will force the adoption of automated taxonomy tools because of the inefficiencies and expenses associated with manual content tagging. Taxonomies, however, provide a representation of the knowledge of an organization and, therefore, their development will never be entirely automated. Automation can help expose and organize this knowledge, but the understanding of a business and its objectives must be integrated into the taxonomy development process. Giga's evaluation criteria will help organizations outline requirements for meeting this balance between automation and manual control. They can also be used to measure specific vendors and products against requirements.

Proof/Notes

Electronic information has grown organically in many organizations, with masses of unstructured content proliferating in a wide range of repositories, from document management and e-mail systems, to the intranet and Web. Some amount of structure is required to manage, access and understand this content better and thus to provide a better understanding of an organization's knowledge resources. Building content taxonomies that can be used to manage and classify unstructured content is becoming increasingly important in many organizations. Accordingly, tools and technologies are emerging to aid organizations in building content taxonomies and classifying documents in a more efficient manner.

The following evaluation criteria can be used as part of the requirements definition and evaluation of taxonomy development and classification technologies. A key theme across these criteria is flexibility. Varying technical approaches to classification can be appropriate in different business situations. There is no one right way to develop and manage taxonomies or one best technology for classifying documents. Overall, look for tools and vendors that offer an array of technical options. Even more important is a useful balance between automation and editorial control.

Taxonomy Generation

Taxonomy creation

Some companies already have extensive content taxonomies in place and may wish to use these in larger classification projects. Others don't have any consistent taxonomies and may desire some automatic generation based on topics found in existing document stores. Tools that support multiple generation options will rank highly in this category. Many products in this segment today require the use of existing taxonomies or the manual creation of taxonomies for the system to use. Look for tools that also enable the import and rationalization of existing taxonomies. Some products provide basic clustering to produce a map of topics

found in various document collections, and this can sometimes be used as a starting point for building a taxonomy. Products that do provide automatic hierarchy generation should be combined with administrative tools to allow for the manual manipulation of categories and editorial input on overall taxonomy structures. A combination of automatic “suggestions” based on existing content and manually created categories to reflect business processes and objectives will likely be most effective.

Taxonomy definition

A variety of technologies can be used to define each node in a taxonomy. This generally is the technology used to classify documents, but it can represent a distinct phase in the overall taxonomy development and classification process. Ideally, a range of options will be available to meet the needs of different taxonomy types and content sources. Possibilities include the following:

- *Rules-based*: These are nodes defined by rules that may be as simple as saying, “All docs in this repository go in this category,” or they may be more complex search strings combining keywords with Boolean queries.
- *Phrase extraction*: Some tools require that the specific categories in the taxonomy be created using noun phrases from documents.
- *Sample/training sets*: Many classification tools today require that each node in a taxonomy be defined using a sample or training set. A classification tool then uses this training set to recognize “topics” based on semantic analysis or patterns-based statistical algorithms. This is the basis for pattern matching in the full document collection.

Taxonomy characteristics

Ensure that a product is able to meet specific taxonomy definition and management requirements. Evaluate the depth of taxonomies it can create, meaning the number of levels it can support. Ensure that the vendor has worked with document collections of an adequate size. It may also be necessary for the taxonomy to support cross-references.

Classification

Classification technology

The technology used to classify documents is akin to the technique used to define taxonomy nodes. If sample or training sets were used, the classification engine will use these sample sets as a basis for identifying patterns and matching documents with similar patterns to specific training sets and taxonomy nodes. Various algorithms are used to do pattern matching. A business rule-driven approach will crawl documents and classify documents based on their ability to meet a rule, within a specified level of confidence or relevancy. When categories are defined by search strings, linguistic tools can often be applied to ensure that queries are expanded to include synonyms, alternative spellings and the like. Tools based on semantic analysis or phrase extraction will determine a document’s “aboutness” based on key phrases.

Obviously, important criterion in this area is accuracy. However, this can be difficult to compare without running pilots of various technologies on specific content stores; thus, this is highly recommended. Again, look for tools offering flexibility so that administrators control how different areas of varying taxonomies can be defined and populated. Manual classifications should also be allowed. As stated earlier, automation must be combined with control. The system should be able to ask for help when it can’t classify documents within a certain measure of confidence. Ideally, the system will learn from manual intervention to make “smarter” categorizations.

Tagging

Classification systems will either build a metadata catalog that provides an index to various document stores, or they may insert tags into documents. A metadata catalog will make taxonomy changes (and retagging) easier. However, placing tags on documents can make the tags more useful and more accessible to different

applications because tags are available as part of a document as opposed to being stored separately in another source. If there is a metadata catalog, ensure there is an exposed API so this catalog can be made accessible to other applications. Tags should be stored in standard Extensible Markup Language (XML) formats.

Document processing

Content needs to be classified in a flexible and scalable way. Initial classifications should be done by batch, but ongoing classifications should be done as part of production processes or via incremental reindexing. Scheduled indexing should also be possible.

Document/data support

Evaluate the variety of document and content types the system can classify. This should include all basic file types, such as MS-type, PDF, HTML, etc., as well as Notes or Exchange content, such as e-mails or discussion content. The system should be able to access and classify structured and unstructured content. Think about the ability to classify not just content, but also people, based on their actions as authors, readers or participants in collaboration forums. Collaborative content should also be classified to include not just a shared document or content in a discussion thread, but also the idea of a “place” based on its overall content and participants.

Multilingual support

The system should be able to classify documents in multiple languages. Ideally, it would also use some cultural knowledge built into the system for common synonyms and a way of editing for local variation on the meaning and interpretation of words. Statistical systems are more likely to be multilingual because they look for patterns that occur in text and have little or no knowledge of the underlying semantics, while keyword and other language-dependent systems are likely to be less adept at languages until expertise is explicitly coded into the system by the vendor or end users. As with all of these tools, the evaluation should include extensive testing of the multilingual features to determine they are adequate for the application.

Maintenance

Administration tools

Administration and maintenance is crucial to the ongoing effectiveness of content taxonomies. Administrators need robust tools that will give them control over the taxonomies and classifications. This should include the ability to override any automation, meaning automatically generated categories or specific document classifications. An administrator should be able to view and manage an automatically or manually generated taxonomy easily. There should be easy-to-use tools to define each node, whether this is by business rule, training set or other means. An administrator should be able to create subcategories or merge categories that may have been manually or automatically created. There should be delegated administration features so that relevant content owners can manage areas of the taxonomy.

Taxonomy maintenance

An administrator should be able to adjust categories manually or add new categories over time, using the administration tools described above. However, the system should also assist in this process by suggesting new categories when it becomes difficult to classify content into existing categories. At a minimum, an administrator should be alerted when documents are found not to fit into any existing categories within a certain degree of confidence. As changes to the taxonomy are made, these changes need to be propagated through content that has already been classified.

Presentation and Access

Presentation options

One of the benefits of instituting content taxonomies is that they can aid in providing improved access to disparate content. Therefore, it is important to evaluate how these taxonomies are presented and exposed. Look for tools that provide a browser-based presentation layer for browsing and navigating a taxonomy. In some cases, taxonomies are exposed “as is” for end-user browsing. In others, taxonomies represent building

blocks providing metadata that can be used to build customizable document directories. Again, flexibility is key.

Visualization tools

Beyond straightforward browser-based access to content taxonomies or document directories, content visualization tools can be a useful complement to classification tools. This means technology that visually represents the breadth and depth of content and relationships between different areas and allows for exploration (see Planning Assumption, [Visualization Tools Key to Exploring Unstructured Content](#), Daniel W. Rasmus).

Search

It is critical that search and taxonomy tools be well integrated. Search should leverage classification so that searches can be narrowed or filtered by different attributes and results can be sorted into categories. If a classification vendor does not offer an integrated search tool, determine how this functionality will be provided since it is crucial for end users. A search engine should be able to provide full-text search across all indexed documents, offer searching of specific metadata fields as provided by a classification tool, sort results by category, narrow results by category or expand searches by category. Combining search with classification allows retrieval to move beyond straightforward keyword searches to concept-based searching since results are driven by the concepts represented in the taxonomy as opposed to just the presence of a particular word in the text of a document. A classification tool that builds a metadata catalog that is external to the documents may need to feed this data to the search engine so that metadata tags are indexed and associated with documents as part of the full-text index. Search of the taxonomy (nodes or category names) should also be available as part of the classification product.

Relevancy

Improved access to information is one of the reasons for developing and implementing content taxonomies. Thus, it is important to determine how the classification engine will rank or display documents that are categorized together. There must be additional relevancy factors to determine document ranking or display order, and these should be tied to content quality. This means sorting search results within a particular category or within the browseable categorical structure. This relevancy sorting should be based on usage data or document age or link analysis, etc. and should be customizable by administrators.

Security

Users should not be able to see links to documents in a browseable hierarchy that they do not have privileges to access. The classification tool has to mirror access control in underlying repositories so that authenticated users only see those documents or even portions of the taxonomy they are entitled to access. Alternatively, the classification tool must integrate with existing authentication and policy management mechanisms to ensure secure access.

Personalization

One step beyond secure access is personalized access. End users should be able to customize their view of the taxonomy or document directory. This may be done through explicit preferences, where users indicate which pieces they want to see regularly, or it may be based on implicit data about users' past usage of content — be sure to offer opt-out capabilities in this approach. Alerting is also an important personalization feature so that end users can subscribe to categories and be alerted when new content is added.

Vendor Evaluation

Vendor viability

Many of the vendors in the taxonomy and classification market are private and small. Vendors like **Lotus**, **Verity** and **Microsoft** are beginning to add classification capabilities to larger product suites, and this will make it more difficult for smaller, classification-only vendors to compete. It is likely in the future that classification technologies will be bundled into other products that attempt to better manage unstructured

content, including Web content management (WCM) systems and enterprise portals. This may make the smaller vendors acquisition targets; there have already been acquisitions in this vein, with **Interwoven**'s acquisition of **MetaCode** last year serving as a good example. Evaluate the financial stability and long-term viability of potential vendors.

Customers

Part of viability is looking at existing customers. It's also important to discuss implementation issues or difficulties with existing customers. The vendor should have some appreciation for domain-specific content issues, especially in highly technical fields, like medicine or engineering. The lack of appreciation for specialized domain content could become an issue when the classification engine encounters phrases it cannot interpret properly, perhaps to the extent of nullifying the value of the classifications. The existence of customers in a domain segment needs to be closely evaluated to see the type of content being classified to determine if customers are using it for horizontal content or specialty content. The existence of specialty content customer should act as an assurance to customers in the same domain.

Partners

The amount of content stored in proprietary systems remains high, which means that most content classification systems will not be able to include content in those stores simply via a standard interface, such as XML or SQL. In these cases, content classification vendors will need to develop partnerships so that their tools sufficiently understand not only the access methods to content stored in repositories, but also how to interface with that content for tagging or metadata catalog access. The existence of partnerships with an organization's primary and secondary repositories should be a strong indicator of fit once functionality and quality criteria are met.

Services

The development of a taxonomy is not a turnkey solution. The development of solid training sets, the modification of the taxonomy and the description of business rules can be daunting to organizations as they deploy classification technology. The existence of a strong services organization or a services partnership should be considered closely during an evaluation. The lack of services will mean lengthy deployment times or failed deployments.

Cost

In terms of software licenses, significant variation exists among the vendors as a raw comparison and in terms of licensing models. Software costs need to be evaluated on a model basis, where each vendor is asked to comment on pricing for a particular model that includes the number of seats, amount of content, frequency of update, number of servers, platform, etc. This will allow each vendor in an evaluation to provide a similar response and zero out any factor that does not contribute to the acquisition or maintenance cost of their software. The total price for acquisition and maintenance will net out in aggregation.

It is extremely important with content classification tools to include all the nonsoftware and hardware-related costs in the model. These include taxonomy building, training set development, end-user and administrator education, rules development, result verification, building special repository interfaces, taxonomy reconciliation and the processes that will support the deployment and maintenance of the system.

Alternative View

In many organizations, simple search tools, like those associated with Web server technology, have not found wide acceptance (see IdeaByte, [GigaWeb Survey Reminds That Site Search Is Often Overlooked](#), Kathleen Hall). Because of this, the addition of search technology may prove sufficient for many organizations during the next 18 to 24 months, which will result in a continued slow adoption of automated classification technology outside special functions, like competitive intelligence and research organizations. Most organizations should concentrate on the deployment of search, the quality of content on their intranets and stronger links between search and collaboration before worrying about any kind of horizontal approach to

automated classification that will increase the costs of content management without a clear return.

Findings & Recommendations

Given the explosive growth of electronic information in most companies today, more efficient means of classifying and accessing information are required. A balance between manually created taxonomies and automatic classification tools will be most effective.

Determine the appropriate balance between automation and editorial control based on the following:

- Taxonomies already in place or in use
- Specific tagging objectives, i.e., knowledge discovery, personalization, etc.
- Availability of taxonomy development skills
- Success of technology pilots on specific content sets

Look for tools that do the following:

- Offer a useful balance of automation and editorial control to match requirements
- Present a variety of taxonomy-generation options, ranging from the import of existing taxonomies through the automated generation of hierarchical categories
- Provide categorization options, including business rules, pattern-recognition technology and the incorporation of knowledge bases and lexicons
- Offer different presentation and integration options so the benefits of taxonomies can be maximized across applications and sites
- Ease maintenance and administration

Vendor selection should include an investigation of the vendor that checks for a customer base, solid funding, business partners, methodologies for deployment, documented internal development processes, patents and services. Services should either be provided by the vendor or delivered via a partner with a track record.

Organizations must evaluate the classification technology and vendors simultaneously. Since this is an emerging market, there are a number of small vendors in this space. Ensure that the vendor has successful implementations and that it is financially viable.

Content classification cannot be automated completely. The most “black box” of systems still requires the selection of documents for training sets and the testing of classification against assumptions to see if classification is adequate. For more sophisticated systems with rules, taxonomy editors and other administrative features, the amount of work will increase as a function of classification accuracy.

As electronic information continues to grow, organizations will demand some amount of structure to manage, access and understand this content better and thus provide a better understanding of an organization’s knowledge resources.

Evaluation criteria for content classification systems should include the following:

Taxonomy generation

- Taxonomy creation
- Taxonomy definition
- Taxonomy characteristics

Classification

- Classification technology
- Tagging
- Document processing
- Document/data support
- Multilingual support

Maintenance

- Administration tools
- Taxonomy maintenance

Presentation and access

- Presentation options
- Visualization tools
- Search
- Relevancy
- Security
- Personalization

Vendor evaluation

- Vendor viability
- Customers
- Partners
- Services
- Cost

References

Related Giga Research

Planning Assumptions

[IT Trends: Knowledge Management — The Next Wave](#), Daniel W. Rasmus

[Visualization Tools Key to Exploring Unstructured Content](#), Daniel W. Rasmus

IdeaBytes

[IT Solutions Demand for 2001-2002: Visibility on Information and Knowledge Management](#), Daniel W. Rasmus

[Understanding Context and Knowledge](#), Daniel W. Rasmus

[Content Tagging Strategies](#), Kathleen Hall