Five Principles of Intelligent Content Management

By Dan Sullivan

Enterprise information portals and other content management applications have many purposes, but most of them share a common goal: to give large numbers of knowledge workers access to large amounts of unstructured text.

EXECUTIVE SUMMARY

Dan Sullivan

As repositories of unstructured text grow, simple search engines will fail to meet all the information retrieval needs of users. Managing content metadata, profiling users, controlling access, supporting rich searching, and automatically gathering content will become essential elements for managing text-based resources.

Content management applications have become increasingly integrated and sophisticated, but nevertheless, we continue to suffer the fact that no single tool, algorithm, or technique can solve the problems posed by enormous volumes of information expressed in complex and ambiguous natural languages. The range of possible topics, our means of expressing them, and the difficulties in teasing out the structural relationships within natural language texts are so demanding that we are unlikely to find a "silver bullet" anytime soon.

In this article, I'll explain how "intelligent" content management techniques that explicitly
represent and exploit implicit information and underlying relationships in unstructured text can help senior IT managers and business leaders give their users better control over the information retrieval process. We can classify these techniques into five design principles that serve the same objective: reduce the effort required to find relevant content such as marketing reports, customer communications, and news in ever-expanding repositories of text.

**PRINCIPLE 1: MAKE METADATA KING**

Metadata about content serves several functions. First and foremost, metadata describes the essential aspects of text, such as main topics, author, language, publication, and revision dates. This type of metadata is designed to improve the precision of full text and keyword searching by letting users specify additional document attributes. It is also helpful for classifying and routing content, purging expired texts, and determining the need for additional processing, such as translation.

Managing content metadata involves two main challenges: extracting it from text, and then storing it. The extraction process needs to detect information ranging from the author's name to the dominant themes in the text while the storage process must efficiently support a number of access and retrieval methods. Manual extraction of metadata is effective on a limited scale, but in general, automated methods such as Solutions-United's MetaMarker, Klarity's eponymous metadata extraction tool, and Inxight Software Inc.'s (a Xerox spin-off) Categorizer are more helpful. Automated methods will not produce the same quality of metadata as a human being would, but they are effective for most purposes.

Regardless of how metadata is generated, you can store it in two ways. First, you can manage it separately from documents in a relational database. This approach is typical in document warehousing, where tight integration with other database applications, such as data warehouses, is required.

Alternatively, you can store the metadata directly within the document. In this case, the XML-based standard Resource Description Format (RDF) is your best bet. RDF is not tied to a particular metadata standard, so you can apply it to a variety of requirements.

Metadata can also enable access control. For example, you could use explicit attributes of a text resource - such as a copyright or license agreement - to manage the distribution of content. For example, corporate libraries that license electronic subscriptions to business and scientific journals might need to track the number of article downloads or restrict access to particular departments or a fixed number of concurrent users. Unlike content metadata, access-control metadata exists at several logical points, such as at the document, journal, magazine, or publisher level.

Quality control and document ranking is another (and often overlooked) use for metadata. Not all texts are created equal, and your information retrieval programs should reflect that fact. For example, you can safely assume an article from *The Wall Street Journal* on foreign investments in Latin America is accurate, but what about a piece from an obscure Web site dedicated to the same topic?

Even internal documents vary in importance. A final report to the COO should have more weight than draft memos circulating among analysts, yet the memos could easily rank higher in a search based simply on word frequencies.

Metadata is not limited to content description, access control, and quality control; you can also extend it to include automatically generated summaries and clustering data. Depending on the
application, however, both summary and clustering information might be generated more effectively on an as-needed basis rather than explicitly stored.

**PRINCIPLE 2: KNOW THE USER**

Representing a user's long-term interests through a profile is yet another key to improving the precision and recall of information retrieval. Profiles, which are explicit representations of these interests, generally use the same representation schemes (such as keyword vectors) as the metadata describing the contents of documents.

Tsvi Kuflik and Peretz Shoval, who research information filtering techniques at Ben-Gurion University in Israel, have identified six different kinds of profiles:

- **User-created profiles** are the easiest to implement but put the burden on the user to create and maintain them.
- **System-generated profiles** analyze word frequencies in relevant documents to identify patterns indicative of interesting texts.
- **System- plus user-created profiles** start with an automatically generated profile that the user can subsequently adjust.
- **Neural-net profiles** are trained using interesting texts provided by the user to output relevancy rankings for other texts.
- **Stereotype models** of interests shared by a large group of users can provide the basis for building individualized profiles.
- **Rule-based filtering** implements explicit if-then rules to categorize content.

Each of these techniques has its benefits and drawbacks, such as requiring manual updating by users when interests change or slow adaptation to such changes. But whichever one you use, the generated profile provides a long-term resource for filtering, disambiguation, and document gathering.

**PRINCIPLE 3: CONTROL ACCESS TO CONTENT**

In contrast to the free-flowing information model of the Web, effective portals require controlled access to content because users are more likely to share information when they know it is distributed within the bounds of well-defined security business rules. In general, content can be grouped into three broad access control areas: open-access information, license-restricted information, and privileged information.

- **Open-access information** is freely available to all portal users. News feeds, press releases, product catalogs, and other publicly available information fall into this category.
- **Access to license-restricted information** is defined by agreements with content providers and applied to significant portions of all available content, such as the entire digital library of a professional association. In these cases, you can adequately control access through basic user authentication or IP address verification.
- **Privileged information** is the most challenging because access is granted on a need-to-know basis. For example, attorneys working on negotiations with one client will need access to related documents, while others working in the same office working on the same type of negotiations but with another firm should not be privy to those same texts. In such situations, the advantages of database-oriented applications become apparent. Relational databases natively provide role and privilege models for access control, and you can implement programmatic procedures for finer grained access control based on content metadata.
When content is adequately protected, you can then turn your attention to creating a navigable repository.

**PRINCIPLE 4: SUPPORT RICH SEARCHING**

Efforts to improve searching have led to a variety of techniques for representing documents, enhancing user queries, and finding correlations among terms. Nevertheless, most information retrieval systems still hit a wall around the 60 to 70 percent range of precision and recall because they depend on primarily statistical techniques instead of linguistic understanding (which is not yet feasible on a general scale). Although we could continue to squeeze marginal gains from keyword searching, a better approach is to combine three techniques: keyword searching, clustering, and visualization.

**SCATTER/GATHER**

The Scatter/Gather algorithm, which was first described by Xerox PARC researchers Douglass Cutting, David Karger, Jan Pedersen, and John W. Tukey in a 1992 paper, uses text clustering to group documents according to the overall similarities in their content. Scatter/gather is so named because it lets the user "scatter" documents into groups, "gather" a subset of these groups, and then rescatter them to form new groups.

The most effective keyword search techniques expand the user's query. A thesaurus automatically adds synonyms to queries, so a search for "stocks" becomes a search for "stocks or equities." Stemmers are used to account for inflections and derivations, so searches for "African" will also find "Africa," and "banks" will check for "bank." Soundex and fuzzy matching are also useful for compensating for misspellings.

But even with relatively high precision and recall, keyword searches can yield seemingly unmanageable number of hits. Clustering is an effective way to address this problem.

Hierarchical clustering is the process of building a tree structure in which the root of the tree contains all documents, internal nodes contain groups of similar documents, and the size of the groups decrease as you move farther from the root until the leaf nodes contain only single documents. These clusters provide a familiar taxonomy-like structure that let users navigate from broad collections of topics to more narrowly focused texts.

Another technique that has proven quite effective in reducing the time it takes users to find relevant content is the scatter/gather algorithm. With scatter/gather, a result set is clustered into
a small, fixed number of groups. (Five seems to be a good size.) Users then select the most relevant of the groups and the documents within that group are then clustered into the same number of groups. Again, the user can drill down into the most relevant cluster and each time the elements are grouped into a number of semantically related clusters. The advantage of this approach is that users can dynamically direct the clustering process as they focus on the most relevant topics.

Clustering effectively groups documents based upon content, but sometimes users need to explore the areas around hyperlinked documents. For example, a member of a geographically distributed sales team might look into sales to a consumer electronics retail chain and find several types of documents ranging from meeting notes of other team members to news feeds from Comtex News, Factiva, or other business content aggregators.

Course-grained navigation tools, such as Inxight's Tree Studio, display hyperbolic trees where each node in the tree represents a document labeled with a title or other descriptive text. Instead of clicking through to each linked page individually, a user can quickly assess a neighborhood of hyperlinked documents and focus in on topics of particular interest.

When users are looking for targeted information, a more fine-grained navigation tool is appropriate. For example, if an account manager is looking for information about sales of mobile phone service and needs to distinguish key marketing terms among a variety of plans, then a tool such as Megaputer Intelligence Inc.'s TextAnalyst allows users to quickly focus on particular terms and discover their relationship to other terms in the text.

Of course, searching, clustering, and navigation all presume a significantly large repository of relevant content. That fact leads us to the final principle.

**PRINCIPLE 5: KEEP CONTENT TIMELY, AUTOMATICALLY**

Some aspects of content management should be automated to keep pace with the available supply of potentially useful content. First of all, you can use harvesters, crawlers, and file retrieval programs to gather documents for inclusion in the content repository. These programs are themselves driven by metadata about which sites to search and which directories or document management systems to scan for relevant content. In many cases, only metadata about documents and indexing detail need to be stored in the portal or document warehouse, and the documents themselves can be retrieved on an as-needed basis.

Automatically gathered documents may require file format or character set conversion before indexing, clustering, metadata tagging, and other text analysis tools can go to work. Automatically managing content is as difficult, or more so, than the extraction, transformation, and load process in data warehousing because the structure, format, and range of topics is more varied. This process will require a series of filters, transformations, and analysis steps as text moves into the content repository.

**HIGH FIVE**

1. Make metadata king.
2. Know the user.
3. Control access to content.
4. Support rich
5. Keep content timely, automatically.

Unlike data warehouses that tend to keep historical data, portal content should be purged. Again, metadata about document types and sources will drive this process. For example, analysts' predictions about earnings reports become irrelevant when an actual earnings report is issued (unless, of course, you want to track the accuracy of the past predictions). In other cases, we might want to keep only the summary of a text, such as a product recall notice or a competitor's press release more than two years old.

Tracking when documents arrived, where they came from, who created them, and other attributes will provide the grist for a number of content management processes.

**REMEMBER THE FIVE**

Free-form text is often called unstructured, but that term is a misnomer. Language's rich structure succinctly represents complex concepts and relationships, but to effectively access that information requires techniques that account for that structure and let users bridge the gap from their interests to information retrieval.

Your organization can realize intelligent content management by adhering to these five basic principles. All are based on the realization that users need to find small amounts of targeted information from sprawling repositories of enormous scale such as the aptly named World Wide Web.