March 2005
http://www.intelligententerprise.com/showArticle.jhtml?articleID=59301538

Structure, Models and Meaning

Is "unstructured" data merely unmodeled?

By Seth Grimes

If managing "unstructured" data is your company's latest must-do — and many people seem to think it should be — take a few minutes to rethink the issue. The claim that 80 percent or more of corporate information is locked in e-mail, documents, audio, images and the like is plausible, but managing it really isn't a significant problem. A variety of content-management solutions handle the job quite nicely.

The imprecise unstructured label and the focus on management divert attention from the real issue: extracting and exploiting the information within binary (as opposed to fielded) data objects. The challenge of modeling and making sense of information content falls in the analytic rather than data management domain.

Most unstructured data is merely unmodeled. Take text, whether written or transcribed from speech. Within the unstructured category, text is of greatest interest to most enterprises. If text didn't have structure, however, documents like this column would be opaque. Text has linguistic structure, both syntactic (grammatical) and semantic (meaning), and texts almost always appear within an envelope of descriptive metainformation such as date, publication and author's name that are used to index documents for storage and retrieval.

Human languages are diverse and irregular, but most humans have all learned to understand at least one, and computers can do the same. Text-mining software applies linguistic analyses and pattern recognition techniques to identify concepts, terms and entities such as names and e-mail addresses. Computers aren't creating structure, they're extracting it by applying linguistic models to documents. We can't expect people to speak computer languages; we want computers to understand our natural languages.

Formal documents have additional narrative or compositional structure: a letter has a salutation, body and signature; a speech has an introduction, arguments and digressions, and a conclusion; and an insurance claim form has a variety of fields. Automated summarization in conjunction with semantic analysis exploits this structure. But even texts that you and your computer can't read contain patterns discernible through statistical analysis. Again, the innovation isn't in structuring text, it's in applying models to discover and exploit their inherent structure.

What about images and video? These data objects also typically come wrapped in sufficient metainformation to support management. And here again, the process of discerning and exploiting patterns is more a matter of creating semantic value than a method of structuring objects. Some text- and image-mining software identify and then tag or extract concepts, entities, terms or
patterns. Some use extracted concepts or patterns to create taxonomies or classification systems to categorize documents. Some apply taxonomies to automate document processing. Some map interdocument links and derive predictive models. These approaches can solve business problems: automating e-mail handling, tuning customer service procedures based on call-center conversations, sifting through medical literature to discover and map disease patterns.

Generic data object management isn't enough because the choice of model (deciding which entities are important, for instance, or picking a taxonomy to classify documents) depends on the business domain. Just as we use different database schemas for different data uses — normalized schemas for transaction processing and dimensional models for data analysis — one size doesn't fit all in document processing. Consider the leading Web search engines: One of their biggest shortcomings is their inability to assess meaning using linguistic and statistical analysis and to provide only the hits the user really wants.

So when analysts and software vendors throw a bunch of disparate problems and technologies into the unstructured data grab bag and try to put the focus on structure and document management, their analyses and solutions are incomplete at best. The spotlight belongs instead on domain-rooted semantics — on automated approaches that discern and apply meaning. I don't expect that we'll stop hearing about unstructured data any time soon, but we must move beyond the hype to understand that analytics rather than data management is the key to getting the most from non-numeric data.

Seth Grimes is a principal of Alta Plana Corp., a Washington-D.C. based consultancy specializing in large-scale analytic computing systems. Write to him at grimes@altaplana.com.

Return to Article