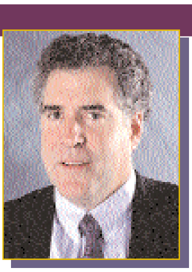
The Problem with Unstructured Data



By Robert Blumberg and Shaku Atre

The management of unstructured data is recognized as one of the major unsolved problems in the information technology (IT) industry, the main reason being that the tools and techniques that have proved so successful transforming structured data into business intelligence and actionable information simply don't work when it comes to unstructured data. New approaches are necessary. This article introduces a series of topics that will look at the emerging solutions for treating unstructured data and some of the business opportunities that this creates. Future articles will cover new search techniques and how search is being integrated with leading Web applications such as e-commerce and self service, classification and discovery systems and the new breed of content intelligence applications that turn unstructured data into valuable business intelligence.

It is well known that one result of the Internet's rapid growth has been a huge increase in the amount of information generated and shared by organizations in almost every industry and sector. Less well known, however, is the degree to which this information explosion has consumed huge amounts of expensive and valuable resources, both human and technical. These demands, in turn, have created an equally huge, but largely unmet, need for tools that can be used to manage what we call *unstructured data*.

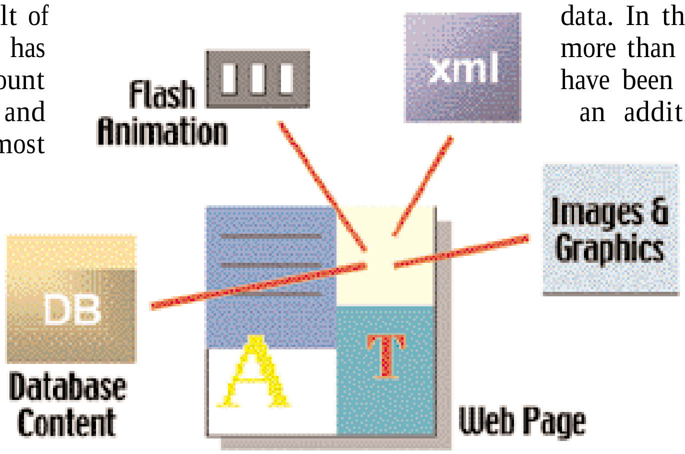


Figure 1: Web Page

The management of unstructured data is a very large problem. According to projections from Gartner, white-collar workers will spend anywhere from 30 to 40 percent of their time this year managing documents, up from 20 percent of their time in 1997. Similarly, Merrill Lynch estimates that more than 85 percent of all business information exists as unstructured data – commonly appearing in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations and Web pages.

Admittedly, the term *unstructured data* can mean different things in different contexts. For example, in the context of relational database systems, it refers to data that can't be stored in rows and columns. This data must instead be stored in a BLOB (binary large object), a catch-all data type available in most relational database management system (DBMS) software. Here, unstructured data is understood to include e-mail files, word-processing text documents, PowerPoint presentations, JPEG and GIF image files, and MPEG video files.

Indeed, a more accurate term for many of these data types might be *semi-structured data* because, with the exception of text documents, the formats of these documents generally conform to a standard that offers the option of meta data. Meta data can include information such as author and time of creation, and it can be stored easily in a relational database management system (RDBMS). That is, while the image or video data cannot fit neatly into relational columns, its meta data can.

One needn't look far to find the source of all this new unstructured data. In the case of the Web alone, more than 2 billion new Web pages have been created since 1995, with an additional 200 million new pages being added every month, according to market-research firm IDC.

The roughly 15 percent of all data that is structured is commonly captured in spreadsheets and databases. In addition, business intelligence (BI) software that lets companies analyze that data in their databases as a way of assisting decision making now proves indispensable for many enterprises.

Why has comparable software for unstructured data – which, after all, accounts for the huge majority of all data in the enterprise – not yet achieved mainstream acceptance?

Raising Awareness

The first step toward solving this enormous problem is raising the awareness of both the users of technology and the companies that design, manufacture and sell it. Awareness is growing, but still has a way to go. In a recent private study, we asked the chief information officers (CIOs) and chief technical officers (CTOs) of 40 major corporations whether they saw opportunities for improved handling of unstructured data within their organization. Of the 40 organizations that participated, 25 (or more than 60 percent) identified unstructured data as a critical issue that could be used to improve operations or create new business opportunities.

Across the board, managers in call centers, technical support and customer-service departments say that while they are generating large volumes of text, they lack ways to analyze the data and thus identify trends and emerging issues. By not having the proper tools in place, they are missing valuable insights. Typical of these managers is an executive at a Fortune 500 telecommunications provider who said, “We have between 50,000 and 100,000 conversations with our customers daily, and I don’t know what was discussed. I can see only the end point – for example, they changed their calling plan. I’m blind to the content of the conversations.”

Another opportunity seen by executives is integration across multiple customer data streams to create broader understanding of customers’ issues. Large organizations that stand to benefit most from content intelligence have their data distributed across multiple data sets, or channels. They also have organizational structures that nearly always result in disparate systems; legacy systems that handle unique, stable, applications; and acquisitions that bring in new systems that need to be integrated. For example, a single division of communications supplier SBC was able to identify *seven* separate sources of customer-interaction data that would need to be integrated to analyze and create a synthetic “total customer voice.”

Location, Location, Location

Once awareness of the issue is raised, the next step is to identify the unstructured data in the organization. In content-management systems, such as those from Interwoven, Web pages are typically considered unstructured data – even though essentially all Web pages are defined by the HTML markup language, which has a rich structure. This is because Web pages also contain links and references to external, often unstructured content such as images, XML files, animations and databases (see Figure 1).

Unstructured data is also prevalent in customer relationship management (CRM) systems, specifically when customer-service representatives and call-center staff create notes. However, once again the verbatim text in call-center and customer-service notes is embedded within a form that is both highly structured and easily represented in a database format.

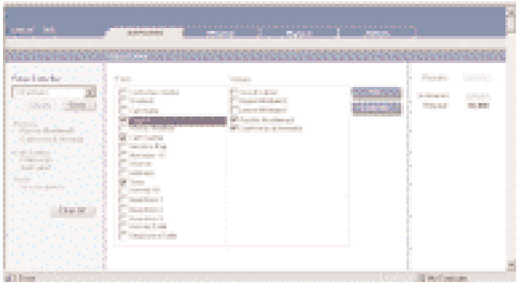


Figure 2: Filter Interface

In sum, unstructured data nearly always occurs within documents. Even though many documents follow a defined format, they may also contain unstructured parts. This is another reason why it’s more accurate to talk about the problem of semi-structured documents.

The Need for Better Searches

A basic requirement for semi-structured documents is that they be searchable. Prior to the emergence of the Web, full-text and other text-search techniques were widely implemented within library, document-management and database management systems. However, with the growth of the Internet, the Web browser quickly became the standard tool for information searching. Indeed, office workers now spend an average of 9.5 hours each week searching, gathering and analyzing information, according to market-research firm Outsell Inc.; and nearly 60 percent of that time, or 5.5 hours a week, is spent on the Internet, at an average cost of $13,182 per worker per year.

Is all this searching efficient? Not really. Current Web search engines operate similarly to traditional information-retrieval systems: They create indexes of keywords within documents and then return a ranked list of documents in response to a user query. Several studies have shown that the average length of search terms used on the public Web is only 1.5 to 2.5 words and that the average search contains efficient Boolean operators (such as *and*, *or* and *not*) fewer than 10 percent of the time. With such short queries and so little use of advanced search techniques, the results are predictably poor. In fact, a performance assessment of the top five Web search engines, conducted by the U.S. National Institute of Standards and Technology, showed that when 2.5 search words are used, only 23 to 30 percent of the first 20 documents returned are actually relevant to the query.

In recognition of the weakness of basic, keyword search, the search-engine vendors have continued to improve their technology. For example, Verity has added techniques such as stemming and spelling correction to its K2 arsenal, while newcomer iPhrase employs natural language processing.

Adding Context to Search

Another problem with Web search engines is that they generally treat each search request independently. This means the results for a given search term will be identical for every user, even when the context differs. For example, if a baseball fan and an amateur bird-watcher both type the words “blue jay” into a search engine, both will receive the same response, regardless of the fact that one is searching for team batting averages while the other seeks a recorded mating song (see Figure 2).

A quantum improvement in search efficiency can be gained by adding *contextual information*. Contextual information generally appears as meta data, and it helps narrow the universe of possible results. One approach, often called "parametric selection," allows users to locate and retrieve information by taking advantage of available meta data by filtering and sorting on known meta data fields such as region, product code or agent name (see Figure 3). By carefully selecting only the relevant fields, the user tremendously narrows the number of records selected for the search.

Another approach is to construct an "advanced search" form that lets users draw on meta data to more thoroughly specify their search. *Nature* magazine's advanced search form, for example, is performed by Atomz's Content Mining Engine on an outsourced or ASP basis. Atomz believes that meta data-assisted search overcomes many of the limitations of standard keyword search.

Beyond Search: Classification and Taxonomy

Structured data that is used for business purposes usually is found in either a spreadsheet or relational database that organizes the data into rows and columns. Similarly, unstructured data can be managed in systems that organize it into a hierarchical structure, referred to as a *taxonomy*. A taxonomy operates like a directory on a PC by providing a convenient, intuitive way to navigate and access information. This means that rather than having to formulate a query and then review the results, the users can instead drill down through the categories and subcategories of the taxonomy until they find the relevant concept or document. The taxonomy can also be used to limit queries to specific categories and sub-categories.

The process of placing unstructured documents within a taxonomy is referred to as *classification*. Often, multiple taxonomies are employed. For example, within a large corporation, two divisions might each maintain separate taxonomies.

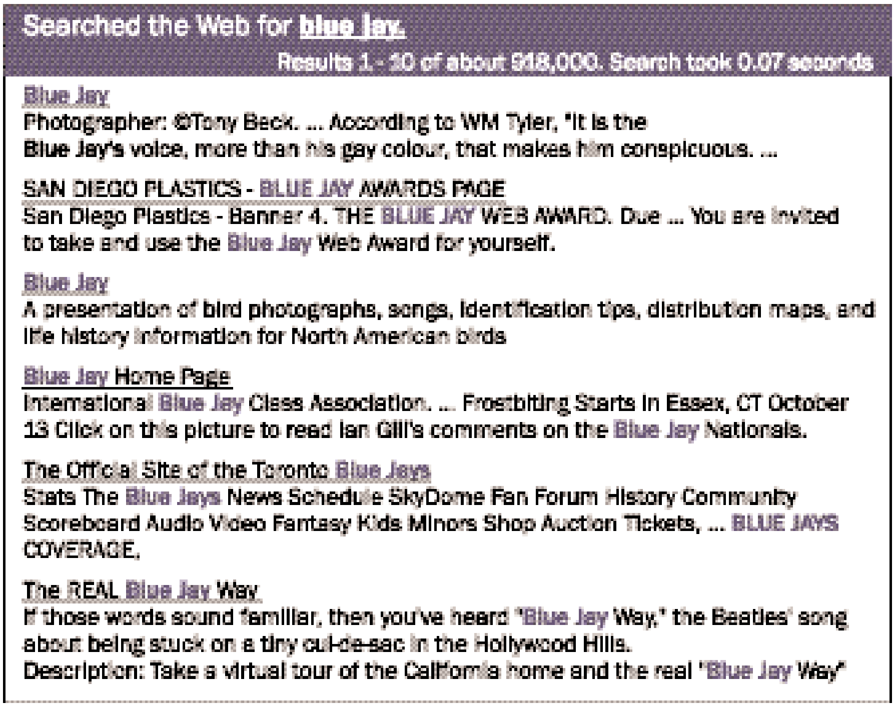
If the industry has settled on taxonomy/classification as the standard way of handling unstructured data, why haven't corporations rushed to embrace this technology? One stumbling block has been the difficulty of creating and maintaining taxonomies. Essentially, this task requires individuals who both understand the organization's business and happen to have a degree in library science – a rare combination. Additionally, a well-populated taxonomy may be eight to 10 levels deep and contain hundreds, even thousands, of categories. To the extent that these taxonomies are based on either product lines or company organizational structures, they need to be changed fairly often. Updating and maintaining such a taxonomy is both time-consuming and expensive.

Fortunately, a new generation of products from companies including Verity, Stratify, Inxight and Autonomy offer automated tools for building and maintaining taxonomies. The first premise behind these products is that documents have subjects or concepts that are important, and the second is that the relationships between those concepts can be analyzed and structured hierarchically. Software that detects these concepts and reveals the relationships among them can build a hierarchy and maintain it over time.

Another major stumbling block has been the unacceptably low accuracy levels of the automated classification systems employed by commercial products. Typically, these products use a single technique, such as Bayesian statistics or a rule-based approach, which works fine in some cases, but not in others. However, several of the latest products are beginning to employ multiple classification algorithms as the vendors recognize that often the algorithm must be selected and tuned to the particular data set being classified.

Content Intelligence: Toward a Solution

It is no coincidence that Documentum and Interwoven, the leaders in content management, both use the term "content intelligence" to describe their search and classification products and services. They, like many vendors, are convinced that with the increasing adoption of enterprise software systems – including content management, CRM and enterprise portals that generate and manage vast collections of semi-structured documents – the next frontier will be value-added intelligent services that generate incremental value for cus-



Searched the Web for **blue jay.**

Results 1 - 10 of about 918,000. Search took 0.07 seconds

Blue Jay
Photographer: ©Tony Beck. ... According to WM Tyler, "It is the
Blue Jay's voice, more than his gay colour, that makes him conspicuous. ...

SAN DIEGO PLASTICS - BLUE JAY AWARDS PAGE
San Diego Plastics - Banner 4. THE BLUE JAY WEB AWARD. Due ... You are invited
to take and use the Blue Jay Web Award for yourself.

Blue Jay
A presentation of bird photographs, songs, identification tips, distribution maps, and
life history information for North American birds

Blue Jay Home Page
International Blue Jay Class Association. ... Frostbiting Starts In Essex, CT October
13 Click on this picture to read Ian Gill's comments on the Blue Jay Nationals.

The Official Site of the Toronto Blue Jays
Stats The Blue Jays News Schedule SkyDome Fan Forum History Community
Scoreboard Audio Video Fantasy Kids Minors Shop Auction Tickets, ... BLUE JAYS
COVERAGE.

The REAL Blue Jay Way
If those words sound familiar, then you've heard "Blue Jay Way," the Beatles' song
about being stuck on a tiny cul-de-sac in the Hollywood Hills.
Description: Take a virtual tour of the California home and the real "Blue Jay Way"

Figure 3: Blue Jay search using popular Web site

tomers. The result is a new generation of enhanced search and classification systems. Major features of these new products include:

Discovery systems: Borrowing a phrase from data mining, content-intelligence vendors are developing interactive applications that help customers explore their unstructured data. Discovery systems, which include Verity's K2, Stratify's Discovery System and Inxight's SmartDiscovery, generate meta data from documents, classify the documents and provide a sophisticated user interface for browsing the document's hierarchy. Each product offers different tools, however. For example, Inxight's product can identify people, places and things, while Stratify's can list related documents.

Platforms for content applications: Leading vendors are developing platforms that expose system functionality through APIs or even XML-based Web services. Such interfaces isolate the application from the underlying information hierarchy, which may change frequently. As a result, users can build targeted applications that apply content-intelligence technology, much the way enterprise applications now rest atop a DBMS.

Data integration: Content-intelligence platforms must let system integrators and customers construct applications that integrate with diverse repositories and legacy systems. As the industry has coalesced around XML as the standard language for exchanging data between systems, vendors have added XML import and export capabilities to their products. Legacy and application-specific formats, such as CRM systems, are handled through specific data import/export modules from companies such as Data Junction.

Integration with enterprise applications: Most types of enterprise software, including CRM, content management and enterprise portals, generate and manage volumes of semi-structured documents. Therefore, it should come as no surprise that leading enterprise software vendors are acquiring and partnering with content-intelligence vendors. A key issue for customers is whether to use the content intelligence systems from their enterprise software vendor – for example, Interwoven Metatagger or Documentum's Content Intelligence service – or to seek a best-of-breed solution from a third party.

Killer Applications for Content Intelligence

Content intelligence is moving beyond search and document classification into full-fledged applications. While the early applications were largely funded by the intelligence community for its own use, commercial applications are now emerging. In fact, "killer applications" are being developed for nearly every industry in which high volumes of semi-structured documents are created. Here are two examples:

Analyzing product defect information for heavy equipment: Manufacturers of expensive equipment such as aircraft and automobiles believe they can minimize warranty repairs by identifying trends within field-service records. One vendor, Attensity, is focusing their unique technology in this area. Says Attensity vice president Perry Mizota, "There is a lot of valuable information tucked away in technicians' notes in field service centers." To get at that value, Attensity's technology transforms text into structured, relational data so that it can be analyzed at a detailed level, much the way structured data is analyzed with BI solutions.

Web-based self service: This solution aims to bring users to a Web site where they can quickly find the answers to their product or service questions. The goal is to reduce the rising costs of providing after-sale service and support by substituting Web visits. According to market studies by Forrester Research, the savings can be substantial: the total cost of a Web visit is $1 to $2, just a fraction of the $50 to $70 cost of the average customer service phone call. To address this need, iPhrase is working with a number of customers, including TD Waterhouse, to integrate their natural language search engine into Web self-service solutions. According to Andre Pino, senior vice president of marketing for iPhrase, "Studies show that 50 percent of Web-site visitors rely on search to find what they need; thus, a complete solution requires that the search engine be capable of interpreting a user's true need and providing an action-oriented response."

The Future Is Now

Content intelligence is maturing into an essential enterprise technology, comparable to the relational database. The technology comes in several flavors, namely: search, classification and discovery. In most cases, however, enterprises will want to integrate this technology with one or more of their existing enterprise systems to derive greater value from the embedded unstructured data. Many organizations have identified high-value, content intelligence-centric applications that can now be constructed using platforms from leading vendors.

What will make content intelligence the next big trend is how this not-so-new set of technologies will be used to uncover new issues and trends and to answer specific business questions, akin to business intelligence. When this happens, unstructured data will be a source of actionable, time-critical business intelligence.

Robert Blumberg, a successful consultant, author, speaker and inventor, is president of Blumberg Consulting. Previously he was president of Fresher Information, a DBMS vendor specializing in unstructured data management, and before that he founded Live Picture where he held various executive positions. He may be reached at rblumberg@inpub.com.

Shaku Atre is President of Atre Group, Inc., a consulting and training firm that helps clients with business intelligence, data warehousing and DBMS projects. She is a former partner with PricewaterhouseCoopers and held a variety of technical and management positions at IBM. Atre is the author of five books including ***Data Base: Strucutred Techniques for Design, Performance and Management*** *and* ***Distributed Databases, Cooperative Processing & Networking****. She is coauthor with Larissa Moss of* ***Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*** *(Addison Wesley, 2003). Atre can be reached at shaku@atre.com.*