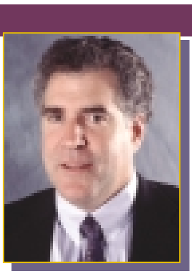
More than Search



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 is the second in a series of articles discussing various aspects of unstructured data.

We are in the midst of an information explosion that is due largely to the rapid growth of the Internet as a tool for conducting business. The proportions of this "infoglut" are indeed epic:

By the end of 2001, the public Internet had become the source of fully half the information used by workers – in excess of 3 billion documents. What's more, according to Google, the Net was doubling in size roughly every eight months.

More than 31 billion e-mail messages were sent worldwide during 2002, according to IDC, of which more than half were sent person-to-person. IDC predicts this number will nearly double by 2006, to exceed 60 billion messages.

One dire consequence of this "infoglut" is that information workers spend ever-increasing amounts of time using corporate networks and the public Internet to search for information. In fact, one study of sales and marketing staff conducted by Outsell Inc. found that employees actually spend more time gathering information from the Internet than they do using it.

Compounding the problem, the majority of Web-based information – more than 80 percent by most estimates – is in a form commonly referred to as *unstructured data* or, more accurately, *semi-structured* documents. These documents include text, graphics, images and movies, as well as Web pages that contain hyperlinked information. In general, data is said to be semi-structured when it cannot be easily searched or processed, unlike the remaining 15 to 20 percent of all data which is structured and, therefore, can be readily absorbed by enterprise databases, business intelligence systems and other enterprise applications.

Tools to search text-based documents, commonly called *full-text search*, have been available for more than 35 years – basically as long as there have been electronic documents. Early uses of the technology focused on a small number of niche, high-value applications, such as legal and intelligence, that have always generated massive quantities of electronic documents. One good example of a business built around these tools is Lexis. Originally founded in 1996 as the Data Corporation, in 1973 Lexis became the first commercial legal-information service to offer full-text search. Today, Lexis offers legal, business information and news from more than 31,000 sources.

Search, as a category of software, has moved well beyond its origins as a function for indexing and retrieving text documents. Today, search software is rapidly evolving into a broad set of tools and techniques for locating products; answering customer-service questions; and exploring, analyzing and processing collections of documents.

Search Software Today

The evolution of search software over the past decade mirrors the development of the Web itself. Originally, both search software and the Internet were tools for locating and accessing information. As the public Web expanded, it enabled companies to both sell (via e-commerce) and support their products electronically. On the enterprise side, a common set of protocols (namely, HTTP/HTML) proved so powerful that it overwhelmed previous efforts to develop a common distributed-computing platform that could be used to link applications, databases and other information sources. In turn, the adoption of these Internet protocols by corporate networks led to the widespread deployment of intranets. This network infrastructure then enabled the emergence of a new generation of enterprise

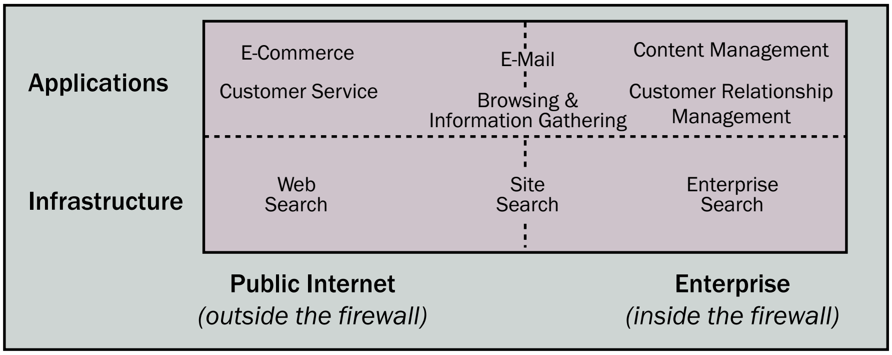


Figure 1: Segmentation of the Search Market

software, including e-mail, portals, customer relationship management (CRM), enterprise resource planning (ERP) and content management (CM).

These enterprise-software suites generate vast repositories of semi-structured documents. In the process, they also create new opportunities for search software at the application level, as illustrated in Figure 1. While a few vendors have focused their efforts on specific applications of search, such as Web self-service, most have instead chosen to offer infrastructure-level products, shown in the bottom half of Figure 1. These products address a broad range of both vertical industries and horizontal applications.

In general, infrastructure-level search products fall into three overlapping but distinct markets:

- **Web Search:** The entire public Web is indexed and made available through a search box on selected Web sites. Leading search sites include Google, Yahoo and Ask Jeeves.
- **Site Search:** All pages of a single Web site are indexed and made available for search from that site. Leading e-tailers, including Amazon and eBay, were among the first to provide high-quality, comprehensive search on every page of their sites.
- **Enterprise Search:** Unifies multiple repositories within an organization. This lets users search and retrieve documents and Web pages via the organization's intranet.

The major vendors for each of these three categories are listed in Figure 2.

Web Search – Ranking the Results

For many users, the main entry point into the Web is a small rectangular search box available on nearly every site's home page. Behind each of these search boxes is an Internet search engine that runs on a Web server and continuously "crawls" the Web (indexes and updates its content). While these search engines use full-text search technology similar to that of early search engines, they continue to add innovations aimed at better serving the needs of users, advertisers and sponsors that drive their business models. Yahoo! Inc.'s announcement, in December 2002 of its intention to purchase Inktomi, one of the most popular Web search engines, underscores the critical importance of search software. As Terry Semel, Yahoo's chairman and CEO, explained, "The addition of Inktomi's search platform adds both control and flexibility to this important business, thus enhancing our ability to create new and more innovative search offerings for consumers and businesses."

Web Search	Site Search	Enterprise Search
AltaVista	*Ask Jeeves*	*Autonomy*
Ask Jeeves	*Atomz*	*Convera*
Google	*Autonomy*	*Inxight*
FAST	*Google*	*Microsoft*
MSN Search	*iPhrase*	*Oracle*
Yahoo!	*Microsoft*	*Stratify*
	Verity	*Verity*

Figure 2: Infrastructure Search Vendors

One question critical to Web search is how to rank equally good matches. For example, a search on the phrase "auction site" at the popular Google Web site yields nearly 2.5 million results. Assuming all the results are relevant, in what order should they be presented to the user? Google's solution, which in its essential form has been accepted by most other vendors, is to rank the Web pages by importance, as assessed by Web users themselves. Google's PageRank method takes into account both the number of visitors to a specific Web page and the number of other sites that are linked to it. In essence, each Web site visit counts as a vote for that site. Those Web pages receiving the highest PageRank appear at the top of Google's results pages.

Google's solution appears to be quite accurate. For example, the previously mentioned Google query for "auction site" delivered eBay, Yahoo and uBid, in that order, as the first three results. These same three sites, in the same order, are the most popular auction sites, according to NetRating Inc., a service that measures and ranks Web site traffic.

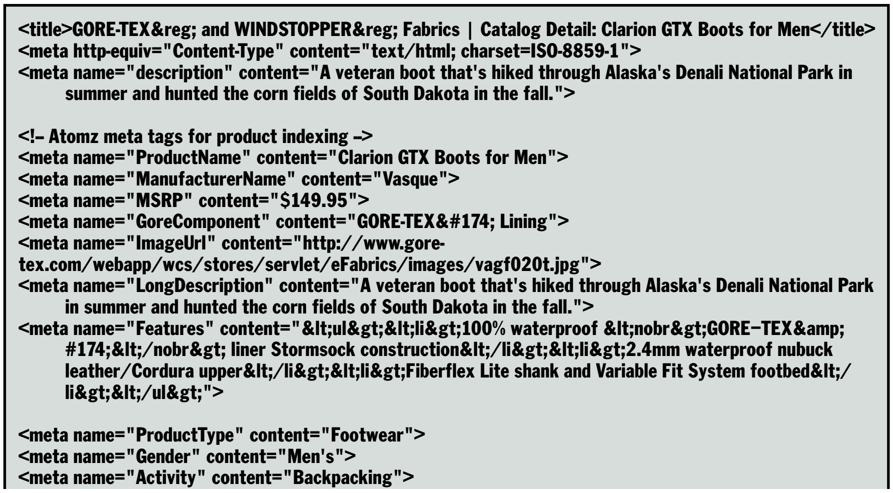
While Google and other Web-search sites equate importance to popularity, site-search and enterprise-search vendors rank their results using very different formulas. Site-search solutions generally try to guide the user to particular pages, products or information consistent with the goals of the site's owner. Enterprise-search engines, on the other hand, attempt to identify the most relevant information or documents for the user, taking into account all available information. Let's examine these two solutions in more detail.

Site Search – Guiding the User

For site search to be effective, it must guide the users to the product or information they seek within just a few clicks. Substantial research shows that the standard ranked list used by Web search sites is among the least effective ways to present search results. Instead, organizing the results into categories of related items has proven to be more effective. In a milestone study, researchers from Microsoft Research and the University of California, Berkeley, evaluated seven alternative formats for presenting search results. The interfaces were divided into two broad categories: list interfaces, which present results as a ranked list similar to the way most Web-search systems work; and category interfaces, which first organize results into hierarchical categories and then rank the results within each of those categories. The Microsoft and U.C. Berkeley researchers concluded that all category interfaces are more effective than all list interfaces.

Not surprisingly, some Web-search sites, including market leaders Ask Jeeves, Yahoo! and Google, have added category information to their standard ranked lists. Some site-search vendors have even taken it up another notch. W.L. Gore & Associates' Gore-Tex site, for example, first resolves searches into product categories, and then presents meta data on as many as four products (including a photo and product name) in a series of rows. If users wish to see more products, they simply click on the appropriate category. For example, a search on "men's boots" returns a series of 11 categories, including backpacking, city wear, cycling, fishing and hunting.

Achieving such precise results requires serious behind-the-scenes work. First, the Gore-Tex Web team defines the products, product information (including product categories) and Web page layout that they want to use. Then this information is provided as HTML meta tags to Atomz, Gore's site-search vendor, which also hosts the search function. Atomz then performs the search and dynamically generates an HTML Web page following the display-format guidelines. To illustrate, the meta tags for the Vasque boot are shown in Figure 4. The words that appear in the "meta name" HTML statements are meta tags and contain explicit meta data that can be used in the search. For example, the "Activity" tag is used to categorize the products into rows. Content that occurs within other HTML tags is implicit meta data and can also be used to search.



```
<title>GORE-TEX&reg; and WINDSTOPPER&reg; Fabrics | Catalog Detail: Clarion GTX Boots for Men</title>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
<meta name="description" content="A veteran boot that's hiked through Alaska's Denali National Park in
        summer and hunted the corn fields of South Dakota in the fall.">

<!-- Atomz meta tags for product indexing -->
<meta name="ProductName" content="Clarion GTX Boots for Men">
<meta name="ManufacturerName" content="Vasque">
<meta name="MSRP" content="$149.95">
<meta name="GoreComponent" content="GORE-TEX&#174; Lining">
<meta name="ImageUrl" content="http://www.gore-
tex.com/webapp/wcs/stores/servlet/eFabrics/images/vagf020t.jpg">
<meta name="LongDescription" content="A veteran boot that's hiked through Alaska's Denali National Park
        in summer and hunted the corn fields of South Dakota in the fall.">
<meta name="Features" content="&lt;ul&gt;&lt;li&gt;100% waterproof &lt;nobr&gt;GORE−TEX&amp;
        #174;&lt;/nobr&gt; liner Stormsock construction&lt;/li&gt;&lt;li&gt;2.4mm waterproof nubuck
        leather/Cordura upper&lt;/li&gt;&lt;li&gt;Fiberflex Lite shank and Variable Fit System footbed&lt;/
        li&gt;&lt;/ul&gt;">

<meta name="ProductType" content="Footwear">
<meta name="Gender" content="Men's">
<meta name="Activity" content="Backpacking">
```

Figure 3: Meta Tags for Gore-Tex Vasque Boot

It would be nice if search techniques like this were adopted for the public Web. However, meta data is applied inconsistently within Web pages, and there are no standard categories. To overcome these limitations, several leading Web-search providers, including Yahoo! and Ask Jeeves, apply their own category structure, or taxonomy, and employ human editors to generate meta data for many Web pages.

Site Search – Cutting to the Chase

Site search is vital – and the consequences of it not working can be catastrophic. In fact, market researcher Jupiter Media Metrix Inc. estimates that 80 percent of online users will abandon a site if the search function does not work well. Similarly, Neiman Marcus Online (NMO) finds that a significant percentage of its online visitors use search to locate products. "Our users are relying on search more than we originally expected," says Michael Crotty, NMO's vice president, "so it's no longer acceptable to deliver an empty search result. By improving search results, we convert more site visits to sales which directly hit the bottom line."

To ensure that happens, NMO selected search vendor iPhrase Technologies Inc.'s One Step product for both the quality of the results and the flexibility to organize and present search results. Like Gore-Tex, NMO displays search results as a series of products with photos and descriptive information, in this case categorized by designer. To help users refine the search quickly, two drop-down menus can be used to select a designer or product category.

While organizing results into categories and using photos and other visual cues both greatly improve usability, the user is still required to make several choices. Yahoo! is working to shorten that process on its popular Yahoo!Finance site. Working with iPhrase, Yahoo! is employing natural language techniques to understand the investor research question and then formulate a tabular response. While natural language interfaces to databases and spreadsheets are not new, iPhrase does an excellent job interpreting the user's intention and providing a direct answer whenever possible. For example, in response to the question "What are the revenues and market caps for software and computer services companies over $20B?" the Yahoo!Finance search function returns the table shown in Figure 4. In this example, the software made a reasoned guess that the phrase "over $20B" meant "companies with market capitalizations in excess of $20 billion." It also accurately guessed that "software and computer services companies" referred to companies in either the software and programming or computer services industries.

Enterprise Search

Enterprise search encompasses a variety of search and retrieval techniques, all of which aim to make an organization's internal use of documents more productive. While site search and Web search are generally associated with the corporate Web site, enterprise search is integral to a company's intranet.

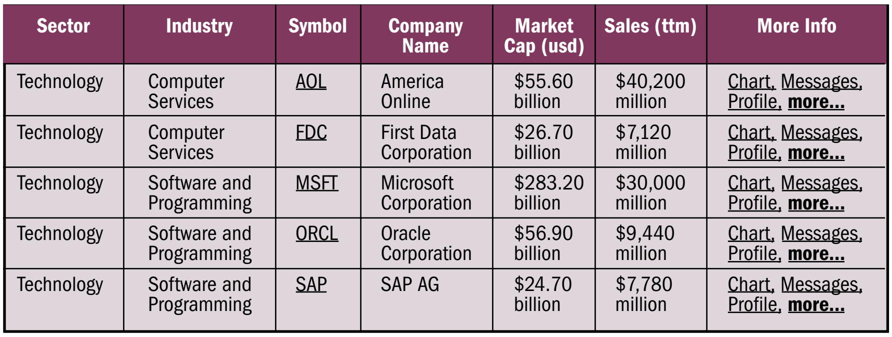
As with site search, enterprise-search usability is crucial. In a recent study of intranets by the Nielsen Norman Group, poor search was identified as the single greatest cause of reduced usability. In fact, Nielsen Norman estimates that a company with 10,000 employees can gain $5 million in productivity each year by improving the overall usability of its intranet, with better search accounting for nearly 45 percent of that potential gain.

In contrast to site search, which often includes basic category information as a way to improve search effectiveness, enterprise search offers sophisticated classification and taxonomy-building tools that integrate search with navigational techniques. Results from an enterprise search typically are first organized into a hierarchical tree of categories, and then ranked by relevance within each category. In addition, most enterprise-search systems also generate a list of similar or related documents.

Enterprise-search systems also retain historical information about the user. This information, together with other information that may be available in a customizable user profile, is used by the search system to determine relevance rankings for a specific user. The goal is to present each user with documents that are closely related to his or her historical interests. For example, if a user searches on "old white house," the system can make an informed guess about whether to deliver results on the President's Washington residence, a historic house or a quaint restaurant.

Crossing the Enterprise

Enterprise-search systems aim to provide robust results so the user won't need to rely on multiple search systems. The systems do this by indexing as many of an organization's documents as is practical. This requires a sophisticated, scalable system that can spider (i.e., locate documents and index them) across different repositories, crossing organizational, geographic, departmental and language boundaries. The system must be architected to operate across a distributed organization with multiple offices and repositories. An overview of the principal functions of an enterprise search system are illustrated in Figure 5. Key features provided by leading-edge systems include:



Sector	Industry	Symbol	Company Name	Market Cap (usd)	Sales (ttm)	More Info
Technology	Computer Services	AOL	America Online	$55.60 billion	$40,200 million	Chart, Messages, Profile, **more...**
Technology	Computer Services	FDC	First Data Corporation	$26.70 billion	$7,120 million	Chart, Messages, Profile, **more...**
Technology	Software and Programming	MSFT	Microsoft Corporation	$283.20 billion	$30,000 million	Chart, Messages, Profile, **more...**
Technology	Software and Programming	ORCL	Oracle Corporation	$56.90 billion	$9,440 million	Chart, Messages, Profile, **more...**
Technology	Software and Programming	SAP	SAP AG	$24.70 billion	$7,780 million	Chart, Messages, Profile, **more...**

Figure 4: Natural Language Search on Yahoo!Finance Site

- **Repositories**: Enterprise search needs to index, search, browse and retrieve documents from a broad variety of repositories. Market-leading vendors Autonomy Corporation and Verity Inc. offer spiders that support leading commercial repositories including Interwoven, Documentum, SAP, Siebel, Lotus Notes, Microsoft Exchange and ODBC-compliant databases.
- **Application Support**: The user will generally encounter enterprise search in the context of an application, the enterprise search function accesses repositories to make content readily available, generally offering the user several ways to retrieve documents including basic search, advanced search and navigation through a hierarchical category structure. Verity's K2 and Autonomy's Portal-In-A-Box search products offer application programming interfaces (APIs) that enable developers to integrate them with their applications.
- **Distributed and Scalable**: Because these repositories can contain thousands, even millions, of documents, it is not practical to duplicate them. Thus, enterprise-search systems create and maintain indices that are kept separate from the physical documents. These indices contain optimized structures to speed the search. They also store meta data for the individual documents. Separating both the indices and the meta data from the documents affords considerable flexibility in constructing distributed and scalable search solutions.
- **Broad Language Support**: All vendors claim to support many languages (English, Spanish, Italian, etc.) and some double-byte languages such as Chinese and Japanese. Autonomy goes one step further by asserting it supports virtually any language. The vendor can make this claim because it uses a purely statistical approach to understanding languages; but even the other vendors perform basic language analysis, including stemming and thesaurus, to provide consistent, high-quality results. Stemming is used to equate words with a common root – such as *run, running* and *ran* – that should be treated equivalently by the search algorithms. A thesaurus is used to equate words with similar meanings (for example, *auto, automobile* and *car*). Inxight, the only enterprise-search vendor whose technology is based on natural language processing, supports 23 languages and offers additional capabilities.
- **Multiple File Formats**: Search must support all document types encountered in an organization. Most enterprise-search products support MS Word, Outlook, PowerPoint, PDF and HTML. Some systems also support media formats including TIFF, JPEG and MPEG, because these include textual meta data fields – such as title, author, date and subject – that can be indexed. Verity claims to support more than 70 different file formats with its KeyView subsystem; most of these formats can be used in either single- or double-byte languages.

The Frontiers of Search

In the past several years, search technology has moved far beyond typing keywords into a box. Today, high-quality results can be reliably achieved using advanced techniques such as meta tagging, classification, natural language processing and historical profiling.

What's more, these technologies – especially those that don't add significant new operations expense but do improve the user experience – are rapidly being adopted by the Web-search and site-search markets.

However, despite the application of this innovative technology, enterprise-search has lagged in the market. Only now is it moving beyond the early-adopter stage of specialized applications such as legal, patent and intelligence to reach the broader corpo-

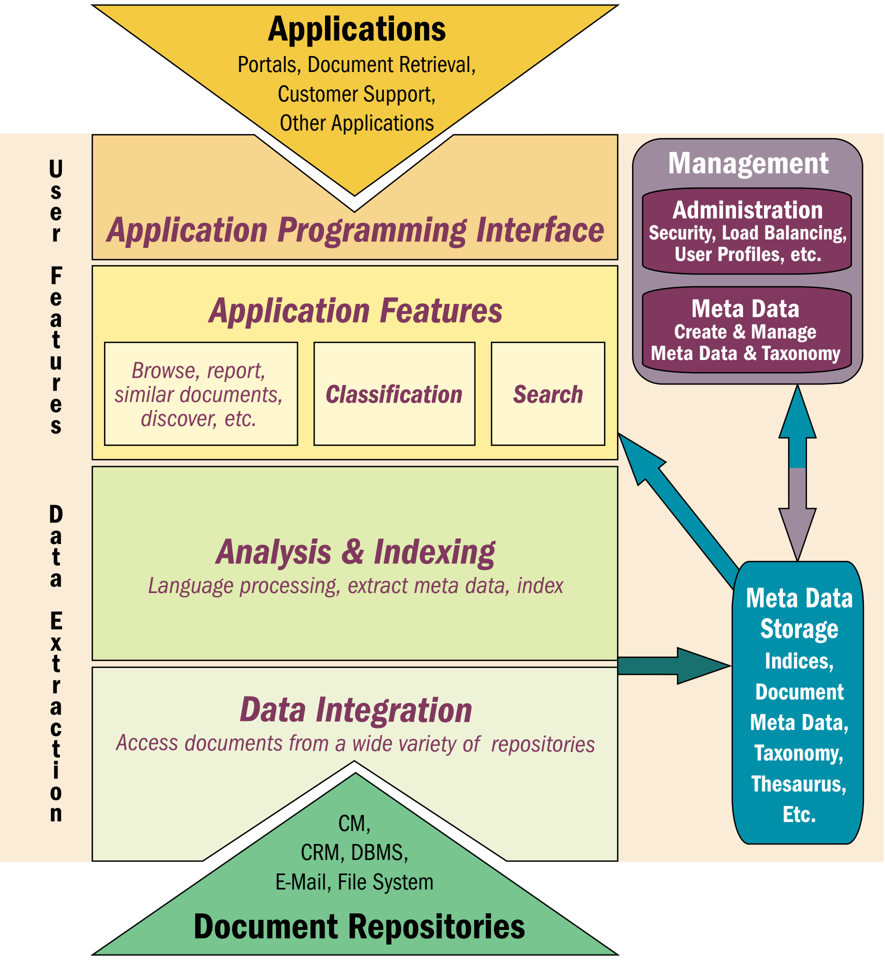

Applications
Portals, Document Retrieval, Customer Support, Other Applications
User Features
Application Programming Interface
Application Features
Browse, report, similar documents, discover, etc.
Classification
Search
Data Extraction
Analysis & Indexing
Language processing, extract meta data, index
Data Integration
Access documents from a wide variety of repositories
CM, CRM, DBMS, E-Mail, File System
Document Repositories
Management
Administration
Security, Load Balancing, User Profiles, etc.
Meta Data
Create & Manage Meta Data & Taxonomy
Meta Data Storage
Indices, Document Meta Data, Taxonomy, Thesaurus, Etc.

Figure 5: Overview of Enterprise Search

rate market. Enterprise search promises to deconstruct the search box and teach enterprises a new search paradigm, one that offers improved worker productivity and other benefits. As Ron Kolb, Autonomy's director of technology strategy, says, "Our business is to get our customers out of the business of searching for data."

This new paradigm, which hinges on classification techniques, isn't exactly brand new, but it does introduce some new capabilities, including *discovery* and *social networks.* These advanced capabilities observe the behavior of individuals in the organization as they sift through and select documents, and then automatically recommend additional subject matter and locate other users with similar interests. The ultimate goal is to first capture and then apply the intelligence inherent in both document repositories and users' actions.

This new generation of search products and technologies can't prevent office workers from being bombarded by an exponentially growing volume of information; however, these products and technologies do promise to transform this infoglut into an information opportunity.

Robert Blumberg is president of Blumberg Consulting, Inc. He has broad experience both as a computer software executive and as a creator of leading-edge Internet technology, products and solutions. Previously he was president of Fresher Information, a DBMS vendor specializing in unstructured data management. Prior to that, Blumberg founded Live Picture where he held various executive positions. Blumberg has been a featured speaker at many industry events and management forums in the U.S. and overseas. 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 of PricewaterhouseCoopers and held a variety of technical and management positions at IBM. Atre is the author of five books, including ***Data Base: Structured 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.*