Discovery:
The Road to Business Intelligence

By Robert Blumberg and Shaku Atre

This is the fifth and final article in a series discussing various aspects of unstructured data.

Previous articles in this series have discussed methods for accessing, navigating and organizing collections of documents. Initially, we discussed how enterprise search has expanded to incorporate techniques that improve the effectiveness of locating and retrieving documents. We followed up with a review of classification techniques that aim to change the paradigm by organizing documents so the user can quickly navigate to the desired information. In this, the final article of our series, we will discuss "discovery systems." These systems are designed to automatically identify important relationships and trends within documents.

Commercial search and classification systems aim at the enterprise market where the widespread penetration of software systems such as e-mail, portals, content management and customer relationship management (CRM) has resulted in an exponentially growing volume of documents within corporate archives and repositories. Techniques that cut through this "infoglut" can tremendously increase employee productivity and efficiency, as the average information worker spends as much as 25 percent of his/her time gathering job-related information, according to recent studies.

Leading enterprise search and classification vendors, including IBM, Verity, Inxight and Stratify, have recently introduced "discovery systems" designed to automatically identify important relationships and trends within documents and document collections. Do these discovery systems introduce a fundamentally new and important set of capabilities that go beyond traditional search and classification? Or are they simply marketing hype aimed at drumming up interest for the latest set of upgrades to existing products?

Before we attempt to answer these questions, let's first explore another: What is discovery, exactly, and when can it be useful? Traditional "information access" techniques, including search and classification, can help users who have a good idea of what they're looking for, says Ramana Rao, chief technology officer of Inxight Technologies. Such users can phrase a specific question or formulate a query. Discovery systems, on the other hand, can help users who have piles of data, but not enough information to phrase a question.

Figure 1: Inxight's Star Tree Taxonomy Navigator

Discovery is an active process. The users interact with the system, either to obtain information about the collection of documents itself or to identify documents that contain important information. In contrast, the principal function of a classification system is to make documents available for browsing, generally through a hierarchical user interface that resembles the Windows file explorer. These Windows-like classification systems guide the users to the most relevant information in the most efficient way possible.

In terms of technology, discovery is directly analogous to the discovery process used in data mining. Both rely on similar techniques, including classification, clustering and visualization. From a product perspective, the recent crop of discovery systems layers new features and capabilities atop automatic text-classification systems (described in our article "Automatic Classification: Moving to the Mainstream," which appeared in the April issue of *DM Review*). The most important added features are the ability to automatically generate and maintain a taxonomy, tools for extracting entities (such as persons, places and things) within the text and visual tools.

Discovery is not a mature application area like word processing in which the feature set is well defined. Rather, it is still emerging. In fact, most discovery systems have been introduced only in the past year or so. For this reason, these systems can be best understood by looking at some of their newer capabilities, especially those that differentiate discovery systems from those used for text classification. Specifically, the capabilities we will explore in this article can:

- Discover relationships between documents or entities.
- Detect trends and issues over time.
- Perform fine-grained extraction of

meta data.
- Visually explore large volumes of data to identify trends and relationships.

Discovering Relationships

The creation of a taxonomy – a hierarchically organized set of categories or concepts from a collection of documents – constitutes perhaps the most fundamental discovery. Rather than dealing with tens or even hundreds of thousands of documents, the taxonomy reduces the top-level information to a smaller number of concepts, categories and entities. It then relies on clustering algorithms (see "Automatic Classification: Moving to the Mainstream," in the April 2003 issue of *DM Review*) to identify the relationships between documents. In this way, taxonomy-generation systems create a hierarchical directory of topics or concepts with which documents can be classified.

For example, Figure 1 shows Inxight Smart Discovery's taxonomy browser. Referred to as a star tree, it lets the user rapidly navigate a complex hierarchical taxonomy or directory. This example depicts how the star tree can be used to quickly browse an extensive news taxonomy. The root of the taxonomy is represented by the small red box at the lower left-hand corner of the screen. In the example, the user has clicked on the "company information" node and is now looking at articles within the "merger and takeover" category.

Figure 3: Historical Trend for "Calling Card" Topic

This ability to structure a new set of documents has a value beyond the initial creation of a taxonomy. New topics, entities, themes and relationships can also be made immediately visible. For example, one user in the intelligence community runs each day's "captured" intelligence (monitored e-mails and voice conversations) through a taxonomy system to get an unbiased view of that day's activity.

Another example, this one from Inxight Smart Discovery (Figure 2), shows the extent to which discovery has moved beyond classification. In response to a search on the word *oil,* the system displays the most relevant documents in its document list window (the large window on the right side of the screen), while the series of smaller windows on the left side of the screen offer the user alternate ways to explore "discovered" relationships. Meanwhile, the "concept filter" window at the top left lists the concepts most closely related to the search term. The user can refine the search by selecting one of these concepts. The category filter window, just below it, displays the most relevant nodes from the predefined taxonomy. In the next window, the "entity filter" lists related entities such as companies, countries and measures.

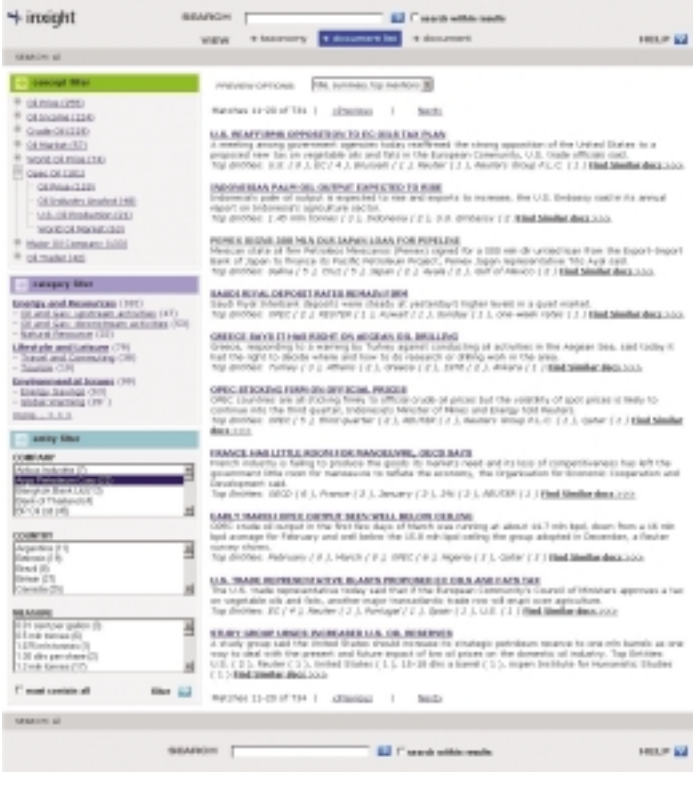

Figure 2: Inxight Smart Discovery's Document List View

Detecting Trends

In practice, a taxonomy acts as directory structure for classifying documents into folders in which the names of the folders are the important concepts and entities are included in the contents of the documents. By maintaining historical values and displaying them using appropriate visual techniques, the system can detect trends. For example, Figure 3 shows the number of calls into a support center that concern the "calling card" topic. A significant increase in volume is apparent starting on April 2, 2003.

Whenever the real world changes, a discovery system's taxonomy must be changed, too. For example, as of March 19, 2003, taxonomies for organizing news needed to add the category: "Operation Iraqi Freedom." Similarly, in the enterprise, taxonomies must be updated to reflect new products, new product names and new operating units.

As a result, an automatic taxonomy system must provide a way to update or modify the taxonomy so that it reflects changes to the contents of documents. This presents an additional opportunity to compare the updated taxonomy to its previous state and determine which key changes were made. Viewing the increasing occurrence of selected topics, or the emergence of new ones, can provide valuable information about trends.

For example, the Stratify Discovery System offers a "refine" feature that can use a set of documents to generate a new taxonomy. This refine operation will not only generate a new taxonomy that accurately reflects the relationships in the new document base, but it will also generate a report that shows how the topics were added or moved.

Figure 4 shows the new taxonomy generated by the Stratify Taxonomy Manager following a refine operation. The example depicts one node of the "Health and Human Safety" taxonomy, before and after the refine operation. The "refined" node has added a new subnode entitled "Occupational

Hazards," which, in turn, contains several subnodes. This significant addition to the taxonomy may have been the result of adding a new source of data such as an external news feed or a new document repository.

For the news agencies and publishers (a significant market segment for discovery systems), this capability presents an opportunity to add products and services. For example, news agencies including Reuters, Factiva and NewsEdge offer narrowband news feeds – that is, a stream of articles pertinent to a particular topic – on a subscription basis. Such companies could also use the refine operation to identify new categories of news, such as occupational hazards, which they then could market and sell.

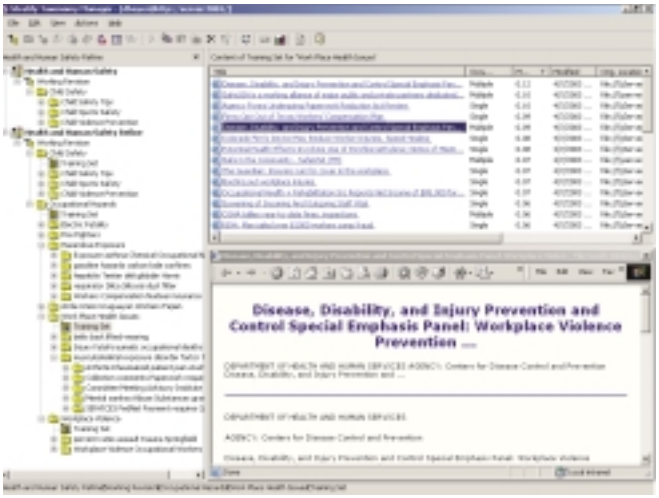

Figure 4: Stratify Discovery System's Taxonomy Manger Following a Refine Operation

Fine-Grained Analysis

This capability of discovery systems has much in common with business intelligence (BI). The market's appetite for BI continues to grow. In Merrill Lynch's December 2002 survey of CIOs, BI software topped the list of anticipated tech spending for this year. Current BI systems, however, operate exclusively on structured data.

Unstructured data, flowing from sources such as field sales forces, e-mails, customer surveys and online forms also contain business intelligence. However, extracting the raw data from unstructured documents requires what Inxight's Rao calls "eyes for text." His company's Smart Discovery product can extract fine-grained meta data from text such as names, dates and locations. These capabilities, referred to as "entity extraction" or "fact extraction," let BI systems exploit unstructured data. Another product with this capability is SAS's Enterprise Text Miner, announced in June 2002. It incorporates Inxight's text technology and exemplifies this new direction by integrating text data with structured data to enrich predictive modeling endeavors.

The payoff for companies that choose to mine their unstructured data for business intelligence can be huge. Applications that have benefited include:

Competitive analysis: Marketing departments use discovery systems to hunt for competitive information among press releases and online user forums. They also identify features in competitors' products that work poorly and are the source of customer complaints. This, in turn, can help both their marketing staff to build highly focused campaigns and their sales staff to pinpoint competitors' weaknesses.

Telecom: Telephone-sales efforts often yield notes that contain references to competitors' pricing and promotions. Sales agents can use discovery systems to sift through call-center notes and distill this information for use in their scripts and Q&As.

Drug and disease research: Pharmaceutical and biotech companies allocate vast budgets for disease and drug research. In the larger companies, hundreds of specialized staff pore over extensive medical and biological archives and repositories looking for links between diseases, trials and research results. Many of these organizations are either implementing, or have already implemented, enterprise search and classification systems to streamline access and retrieval of this information. The new features offered by discovery systems will let them analyze medical research collections to answer questions such as, "Which genes appear most often in the literature relating to a specific disease?" or "Which diseases are most often implicated relative to the mutation of a specific gene?"

Homeland security: Search, classification and now discovery have become essential tools for intelligence and law-enforcement organizations. These groups receive a constant flow of messages and information from field personnel, tips, monitored conversations and e-mails, and intra-agency sources. Nearly every vendor reports that a significant portion of their sales derive from homeland security-related initiatives. Last October's Washington, D.C.-area sniper case provides a good example. The shootings remained front-page news for roughly three weeks, during which repeated pleas for information by local officials and the establishment of a special public FBI hotline, coupled with a $500,000 reward, resulted in the FBI receiving an avalanche of phone calls, letters and e-mails. In fact, FBI tip lines were receiving as many as 15,000 calls per day on 75 phone lines during the investigation. This would appear to be a perfect application for a discovery system that is capable of extracting and then correlating names, locations, vehicle descriptions and license-plate numbers with existing databases and information sources.

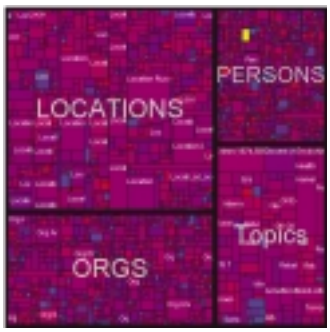

Figure 5: Stratify Heat Map Depicting an Item's Recency and Frequency

Visual Methods

One area in which discovery systems exceed the capabilities of current search and classification systems is in visual methods for exploring data. Most classification systems use navigation through a hierarchical list-like structure as their primary user-

interaction technique. This has the advantage of using a familiar interaction technique – Windows. However, visual methods, many of which trace their lineage to John W. Tukey's seminal book, *Exploratory Data Analysis,* equip the information analyst with powerful tools. These tools can reveal underlying structures in the data and evoke new, possibly unsuspected, insights.

For example, Stratify's Discovery System includes a "heat map" (see Figure 5), that depicts both entities (locations, organizations and persons) and topics in a single rectangular map. Each item is depicted as a rectangle whose size and color represents a different, configurable property. In this example, the size of an item is governed by the number of its occurrences. Its color, similarly, is determined by its relative "recency," with red being the most recent. Using this map, the analyst can identify, at a glance, recent occurrences of a topic or entity. He can then click on an item to zoom in to its "neighborhood" to obtain the next level of information. The heat map lets the user first spot significant occurrences over a very large data set and then drill down for more details.

Another example comes from Cedar Enterprise Solutions, Inc., whose VisuAlert product employs a self-organizing map (SOM) or "thermographic" technique to help customers visualize knowledge (Figure 6).

VisuAlert first extracts from each document (whether e-mail, PDF, Web page or scanned document) its dominant topics or "themes." Then, a neural network calculates the unique location of each document. It bases this calculation on the detected issues within both the document's content and the interrelationship between each of the domains and documents in the knowledge topography.

In the example, the knowledge visualization or SOM is created automatically from the free-form notes within more than 100,000 turbine repair field-service reports from an engineering services company. Black lines distinguish the boundaries of the key issues within each domain. The records that contain a strong relationship with the selected theme – in this case, "system" – are highlighted in blue. The user can personalize the color and shading of the contours and points based on meta data such as file age, number of document reads and author. This will sharpen the focus on the user's key issues. For example, by tracking the growth of "hot spots" over time, a repair supervisor might first determine that rust is becoming a problem in a number of pump subsystems and then initiate a proactive maintenance procedure to reduce expensive downtime.

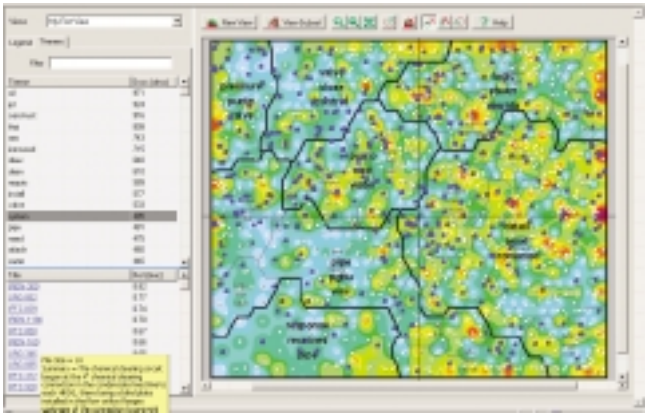

Figure 6: Cedar's SOM –Turbine Repair Example

While heat maps and SOMs will not appeal to all users, they do let managers and analysts rapidly review large, unfiltered data sets and then detect the most important information. In fact, visual tools are most effective for users who, though familiar with their data environment, are struggling to extract the critical issues from a large volume of constantly changing sources, explains Patrick Hogan, Cedar's VP of content value management. "They notice changes quickly," Hogan says. "Visualization lets them identify trends and issues earlier than other techniques." By contrast, he adds, reports or list-style interfaces alone do not reveal the complete perspective, meaning key issues can go undetected.

On The Horizon

While discovery systems targeted at unstructured data have only recently emerged, they show great promise in several directions. In the realm of information access and retrieval, discovery adds valuable new features, including entity extraction and visualization. In this dimension, discovery is poised to augment search and classification systems as a critical "knowledge infrastructure" underlying Web sites, portals, CRM, document-management and content-management systems. In the long run, however, discovery may not emerge as a category independent of classification systems or "enterprise search." Instead, it may simply foreshadow the next round of feature extensions.

Similarly, discovery systems for unstructured data add fine-grained analysis capabilities that complement existing data mining systems. In this regard, discovery may become a bridge that enables BI systems to broaden out and address unstructured data.

Finally, discovery emerges as a new and unique technology area when it forms the nucleus of new applications for drug research, the analysis of repair data, and homeland security. In these and other areas, discovery can offer new insights that yield tangible benefits.

Robert Blumberg is the president of Blumberg Consulting, Inc. He has broad experience both as a computer software executive and as a consultant focusing on product R&D, marketing and business development. Blumberg specializes in the commercialization of software products, from market analysis to product design and go-to-market planning. Previously, he was president of Fresher Information, a DBMS vendor specializing in unstructured data management. Before that, Blumberg founded Live Picture, where he held various executive positions. Blumberg has also been a featured speaker at many industry events and management forums in the U.S. and elsewhere. He may be reached at rblumberg@earthlink.net.

Shaku Atre is the president of Atre Group, Inc., a consulting and training firm in Santa Cruz, Calif., that helps clients with business intelligence, data warehousing and DBMS projects. Formerly, she was a partner at 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 most recently 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.*