Automatic Classification:
Moving to the Mainstream

By Robert Blumberg and Shaku Atre

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

As we've discussed in our previous articles, exponential growth in the volume of documents within corporate archives and repositories has resulted from the widespread adoption of the Internet; the consequent rise of enterprise software systems such as e-mail, portals and content management; and customer relationship management (CRM). This condition is commonly referred to as "infoglut" or "information overload." Most of the information causing infoglut is in the form of documents that contain text, images or other unstructured data that cannot be readily stored in relational databases. We refer to these documents as "semistructured."

As depicted in Figure 1, a number of "content intelligence" techniques are available to work with semistructured documents. The last two techniques in the list, basic search and advanced search (both discussed in previous articles of this series), have been broadly integrated into enterprise applications.

Classification systems are the next frontier because just as a schema structure represents the contents of a database and enables queries, an information hierarchy or taxonomy organizes a repository of semistructured documents. This lets the user navigate a hierarchical structure of categories analogous to the folder systems of Microsoft Windows and the Macintosh OS, which research has shown to be highly effective for locating information.

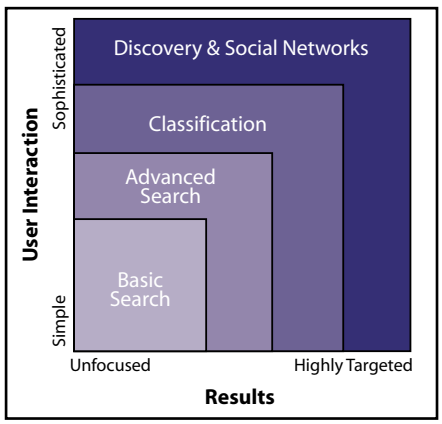

Figure 1: Content Intelligence

Classification has long been used in libraries as a way to organize books, periodicals and other texts, as well as for organizing technical collections. However, in 1995 classification leapt into public view. That year, Yahoo! introduced a Web site, later known as a Web portal, which organized a broad variety of information into categories. Soon, portals covering every conceivable subject area appeared. Yahoo!'s directory of Web sites, unlike many of the other portals, spans thousands of categories and employs teams of human editors who manually classify news and information. For most organizations, this manual approach is economically unfeasible. This, in turn, has stimulated a flurry of research into automatic classification methods. This research has resulted in a new generation of technologies and products, and some of the leading vendors offering these products are identified in Figure 2.

Yet the question remains: Will this new generation of products be able to carry automatic classification into the mainstream enterprise market? Before we attempt to answer this question, let's first look at the latest technology. We'll then examine several representative applications as well as the tools used to construct them.

From Words to Concepts

The goal of all text classification is to assign documents into one or more content categories. While categories are generally predefined, they may be automatically generated based on the content. Any type of document containing text can be classified, including traditional documents such as reports and memos as well as e-mails, Web pages, call-center notes and other less traditional types. Classification is either performed on a document repository, such as a library, or operated on a stream of incoming documents, such as that which might arrive from a news agency or field sales staff.

- Autonomy
- ClearForest
- Documentum
- Interwoven
- Inxight
- Microsoft
- Mohomine
- Open Text
- Oracle
- Stratify
- Verity

Figure 2: Vendors of Automatic Classification Systems

On the technology front in the past few years, vendors have brought the ability to extract and classify concepts, rather than words. This has required considerable language processing. As illustrated in Figure 3, words are first *stemmed*; that is, they are reduced to their root form. Next, *stop words* are eliminated. These include words such as *a, an, in,* and *the* – words that add little semantic information. Then, words with similar meanings are equated using a thesaurus. In the

example, the words *IBM, Big Blue,* and *International Business Machines* are treated as equivalent.

Finally, the classifying tool will use statistical or language-processing techniques to identify noun phrases, or *concepts,* such as "Polaris missile" or "red bicycle." In the example, six noun phrases are identified. Further, using a thesaurus or lexicon, these noun phrases are reduced to three distinct concepts that will be associated with the document. In the example, there are three instances of IBM, two instances of acquisition and one instance of Widget, Inc.

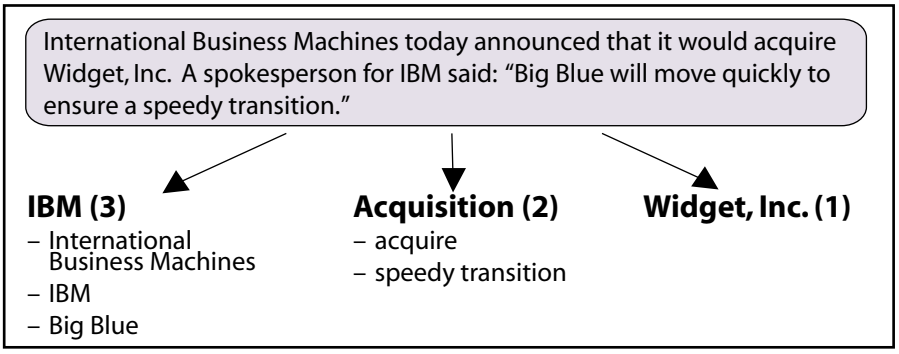

Figure 3: Classification Example

Approaches to Classification

As illustrated in Figure 4, there are four main approaches to classification:

Manual: Often used in library and technical collections as well as in call centers and forms-processing environments, manual classification requires individuals to assign each document to one or more categories. These individuals are usually domain experts who are thoroughly versed in the category structure or taxonomy being used. Manual classification can achieve a high degree of accuracy – although even domain experts will occasionally disagree on how to categorize a document. However, manual classification is more labor-intensive and therefore more costly than automated techniques.

Rule-Based: In this form of classification, keywords or Boolean expressions are used to categorize a document. This is typically used when a few words can adequately describe a category. For example, if a collection of medical papers is to be classified according to a disease, then a medical thesaurus that lists each disease together with its scientific, common and alternative names can be used to define the keywords for each category.

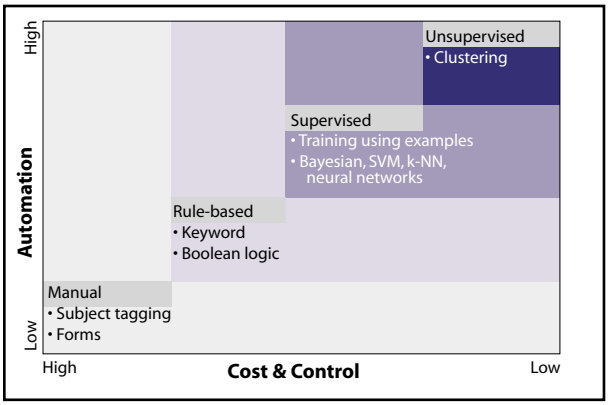

Figure 4: Text Classification Approaches

Another example involves e-mail systems which typically provide rule-based methods for routing messages to specific mailboxes. The e-mails are routed based either on the sender's name or the occurrence of specific words in the subject line. For example, the occurrence of the word "remove" would cause the sender's name to be dropped from an e-mail list.

While rule-based approaches are effective for carefully tuning a limited number of categories, the expense of defining and maintaining categories is generally prohibitive for large-scale classification systems.

Supervised Learning: Most approaches to automated text classification require a human subject-expert to initiate the learning process by manually classifying or assigning a number of "training documents" to each category. This classification system first analyzes the statistical occurrences of each concept in the example documents and then constructs a model or "classifier" for each category that is used to classify subsequent documents automatically. The system refines its model, in a sense "learning" the categories as new documents are processed.

Each vendor has a favored algorithm. For example, Autonomy and Stratify favor Bayesian algorithms, while Inxight and Verity use a form of the k-nearest neighbor (kNN) approach. Microsoft uses support vector machines (SVM) in its SharePoint portal and other products.

Both vendors and researchers have recently undertaken a number of studies to evaluate the performance of various algorithms in different situations. While the studies point out some of the strengths and weaknesses inherent in the various approaches, they also highlight the importance of testing a corpus of documents that is representative of the actual documents an organization wants to handle. Criteria that need to be tested include performance, efficiency (the percentage of documents correctly classified), errors and left-outs (the percentage of documents that aren't assigned to any category).

Unsupervised Learning: These systems identify both groups, or clusters, of related documents as well as the relationships between these clusters. Commonly referred to as *clustering,* this approach eliminates the need for training sets because it does not require a pre-existing taxonomy or category structure. As illustrated in Figure 5, the Autonomy Corporation cluster map tool groups all articles concerning the Venezuelan oil strike in a collection of news articles. However, clustering algorithms are not always good at selecting categories that are intuitive to human users. For this reason, clustering generally works hand-in-hand with the previously described supervised learning techniques.

Which Approach is Best?

Each of the four approaches is optimal for a different situation. As a result, classification vendors are quickly moving to support multiple

methods. As Stratify chief technology officer Ramana Venkata explains, "No single technique outperforms the others in all situations since categories vary in how 'precisely' or 'diffusely' they can be described. Therefore, we need to provide different approaches that can be used simultaneously and an appropriate way to combine the results so that users can achieve optimum results."

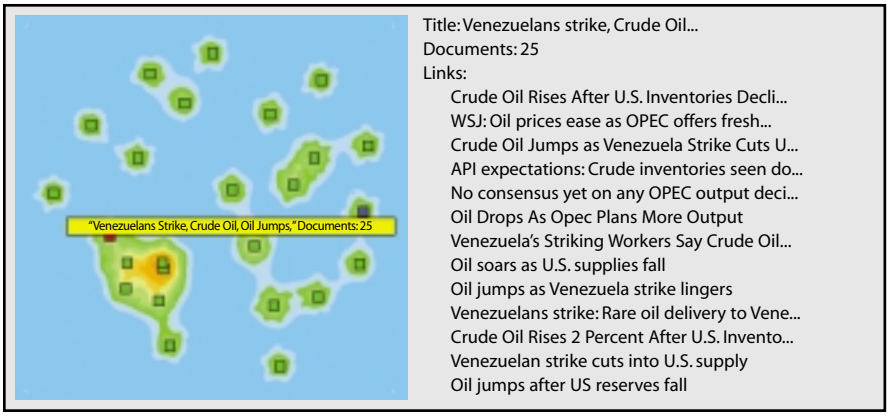

Figure 5: Autonomy Cluster Map

Classification in Practice

Thus far, we've discussed state-of-the-art technology; but what about the state of the market? Most real-world implementations combine search, classification and other techniques such as identifying similar documents to provide a complete information retrieval solution. Following are several examples.

A Research Portal

Cancer.gov (www.cancer.gov), the National Cancer Institute's portal for government cancer research, aims to be the definitive source for research on cancer. The portal provides access to more than 200,000 documents located on roughly 150 Web servers throughout the National Cancer Institute (NCI). According to its project manager, Josh Stella, Cancer.gov has two main goals: making the site easily searchable and navigable for patients, and providing the tools for health professionals to find specialized research. "It's important to include the patients," Stella says, "since they use this as a way of augmenting information provided by their doctors."

When NCI's content specialists process new documents, they review meta data provided by the author, such as author name, title, abstract and keywords. Then they classify the article using UMLS (Unified Medical Lexigraphic System), a comprehensive medical thesaurus containing more than 2 million terms. However, the Cancer.gov site exposes a much smaller category structure through its Web site – in the order of 100 nodes – so that it can be easily navigated by specialists and non-specialists alike.

In building this system, Cancer.gov looked for a product that could combine sophisticated search capabilities such as *spidering* across many repositories (that is, locating documents, extracting meta data and indexing them), load balancing and distribution. Ultimately, Cancer.gov selected Verity's K2 Enterprise search product. It has since developed a highly customized application using K2's comprehensive application programming interface. Project manager Stella notes that Verity K2 was one of the few products that would let him implement "best bets," in which the results of queries that contain common search terms such as "breast cancer" or "colon cancer" would return a preselected list of documents, followed by the actual search results.

Automating Customer Support

Cadence Design Systems, a leading provider of electronic design technology and services, has moved aggressively to provide Web-based customer support. The company has done so by integrating its existing Siebel CRM system with its Web-based customer support portal, SourceLink. On the back end, support engineers answer questions and resolve issues using the Siebel system, regardless of which contact channel the problem enters from – call, Web, e-mail, etc. On the Web side, SourceLink (www.sourcelink.cadence.com) provides application notes, FAQs and other information, and enables customers to enter service requests, receive case numbers and track the case.

Because Cadence uses Interwoven's TeamSite content management system to manage its Web site, it was a natural step to implement Interwoven's MetaTagger product to create a unified meta data repository that would integrate the SourceLink and Siebel systems. MetaTagger extracts meta data, including product name, release, type of document and modification date, from the Siebel documents and then automatically classifies the documents. The user can then browse documents according to a simple three-level taxonomy: product, release, type of document.

The meta data repository created and managed by MetaTagger is also used for document searches by the Verity K2 search engine used by SourceLink. It further enables Cadence to respond quickly to product changes. For example, changing key product information, such as name or release number, can be accomplished within MetaTagger without touching the documents themselves. "Our goal has been to streamline the cycle of publishing a change or introducing a new product onto the Web site," says Vianne Sha, the SourceLink knowledge architect. "With Sourcelink, we've shortened the cycle from months to days."

Narrowband Filtering of News

Factiva, a joint venture of Dow Jones and Reuters, is tackling the infoglut head on. Drawing from 8,000 sources in 118 countries, including the combined content of the Dow Jones and Reuters interactive information services, last year Factiva processed 52 million articles per month to deliver streams of highly targeted business information to the

desktops of its 1.5 million subscribers. From the start, Factiva recognized that neither the manual approach to classifying and assigning meta data that it inherited from Reuters nor the rules-based approach used by Dow Jones would scale sufficiently. Therefore, Factiva licensed Inxight's categorization solution to automatically extract meta data and classify text according to Factiva's extensive taxonomy. "In the three years subsequent to the implementation of Inxight Classification technology, the volume of content that Factiva needs to categorize has increased between 40 and 50 percent, while costs have remained constant," said Simon Alterman, vice president and director of content for Factiva. "This is because Inxight's technology has enabled us to redeploy our human resources into higher value roles – such as customer and product enhancement support – by automating labor-intensive tasks."

Case studies such as those of Cancer.gov, Cadence and Factiva give us confidence that the new generation of classification technology is ready for broad deployment. However, implementers still must face a potentially thorny issue early on – that of taxonomy.

Taxonomy: Cost and Opportunity

Classification systems require a predefined arrangement or division of topics or categories – "buckets" into which documents will be placed. While taxonomies need not be hierarchical, in practice they usually resemble a Windows or Macintosh directory structure with both higher-level groups (taxon) and subgroups (taxa). Perhaps the earliest and best-known taxonomy was published by Carolus Linnaeus, the Swedish botanist, in 1735. The Linnaean system for classifying all living things was widely accepted by the early 19th century and is still the basic framework for all taxonomy in the biological sciences. Formal taxonomies now abound in the natural, social and computer sciences. A few are described in the accompanying sidebar.

Building an extensive custom taxonomy can be a large expense, requiring in-house domain experts as well as specialized consultants. However, a large, sophisticated taxonomy is often unnecessary; and in the cases where it is needed, new tools can reduce the development and maintenance costs. Here are two further tips and tricks:

- Start small: Ron Kolb, director of technology strategy for Autonomy, a leader in content intelligence systems for portals, suggests that many intranet portals will require 30 categories at most. A typical corporate hierarchy might include categories such as products, services, engineering, support, HR policies and benefits. A relatively flat category structure for browsing, augmented with a powerful search capability, will often meet an organization's needs.
- Use existing taxonomies: Organizations in technical areas such as medicine and biotechnology can benefit from existing taxonomies. A common strategy is to start with a taxonomy such as the UMLS, then flatten it in places and add terms specific to the organization in others. For example, ExxonMobil now licenses a taxonomy for equipment classification and performance indicators that it originally developed to streamline internal operations.

Still, organizations that have document repositories they want to share within the organization or with customers will generally benefit from a customized taxonomy. These organizations should look to the most recent classification/taxonomy systems from leading vendors, including Autonomy, Inxight, Verity and Stratify. Starting with a representative corpus of documents, these systems use clustering algorithms to analyze and determine groups of related documents. Clustering will often expose useful relationships and themes implicit in the collection that might be missed by a purely manual process.

Once the documents are clustered, an administrator can first rearrange, expand or collapse the auto-suggested clusters or categories, and then give them intuitive names. The documents in the clusters serve as initial training sets for supervised-learning algorithms that will be used subsequently to refine the categories. The end result is an information hierarchy and a set of topic models fully customized for the enterprise's needs.

Organizations with document collections that span complex subject areas – such as medicine, biotechnology, oil and gas, and aerospace – will inevitably have a large taxonomy. However, there are ways to refine the taxonomy gradually so it doesn't become an overwhelming task. "You can start by importing an existing taxonomy or using supervised learning to create an initial taxonomy," says Ramana Rao, Inxight's CTO. "Then, over time you can refine portions of the taxonomy that are frequently used or contain high-value documents by

Formal Taxonomies

Taxonomies, thesauri and other forms of controlled vocabulary are defined and maintained by many government agencies and industry associations. These taxonomies, originally formulated with the objective of standardizing terminology within an industry or technical subject area, are increasingly used by computer search and classification systems. Here are several examples:

Dewey Decimal Classification system (DDC): Now published by the Online Computer Library Center, the DDC was conceived by Melvil Dewey in 1873 and is today the most widely used library classification system in the world. It organizes materials into 10 broad categories – General, Philosophy, Religion, etc. – and then divides these further into subcategories. While the DDC is still used by 95 percent of the public and school libraries in the U.S., most academic collections have been transferred to the more recent Library of Congress classification system. (http://www.oclc.org)

Medical Subject Headings (MeSH): This system comprises the National Library of Medicine's controlled vocabulary used for indexing articles, cataloging books and other documents, and searching MeSH-indexed databases. (www.nlm.nih.gov/mesh/filelist.html)

The Germplasm Resources Information Network (GRIN): This system for taxonomic data, maintained by the U.S. Department of Agriculture, provides the structure and nomenclature for the accessions of the National Plant Germplasm System. Many plants – some 37,000 taxa and 14,000 genera – are included in the GRIN taxonomy. (www.ars-grin.gov/npgs/tax/)

What are Oracle and Microsoft Doing?

Although neither Oracle nor Microsoft offers an independent classification product; both vendors are actively developing and using classification technology in their products and solutions. Here's a snapshot:

Oracle: According to Oracle's director of product management, Sandeepan Banerjee, "In the same way that structured data has yielded its value through business intelligence, OLAP and transactional systems, solutions involving unstructured data constitute the next frontier for Oracle." Oracle sees classification search, clustering and visualization as features of a comprehensive data management platform, Oracle9*i* Database, that can treat structured as well as semistructured data. In the Oracle9*i* Database, these features are referred to as Oracle Text. Oracle's sweet spot is large-scale search and classification systems. For example, their customers include Online Computer Library Center (www.oclc.org) that connects the cataloging and loan systems of its 43,500 member libraries in 86 countries, and Der Spiegel (www.spiegel.de), one of Europe's most popular magazines, which uses Oracle9*i* Database to store, search and classify documents in its research archive that receives as many as 50,000 new documents per week, pulled from more than 300 publications in 15 languages.

Microsoft: Microsoft's Portal Business Unit has developed sophisticated content intelligence features for inclusion in the SharePoint Portal Server which it selectively incorporates into other Microsoft products, including Office, Windows and SQL Server. According to SharePoint product manager Kyle Peltonen, their goal is to "solve the problem of taking the full breadth of corporate information and making it available through a portal." To this end, Microsoft has focused development on supervised learning technology, using the support vector machine (SVM) algorithm, which they believe is most appropriate for their target market.

improving the training documents or adding specific rules and keywords. The objective is to spend time in areas that yield the greatest payoff. This way, the taxonomy gradually evolves consistently with the directions and needs of your organization."

Headed for the Mainstream

Our case studies confirm that the new technology for automatic classification is both available today and being implemented in a wide variety of applications, including content management, document retrieval, customer service and support portals, and online publishing. Clearly, classification has moved well beyond its early uses in libraries and technical collections into the mainstream enterprise software market. Automatic classification will increasingly be used to access, analyze and make available repositories of information that have been growing and collecting dust – or that have been accessible to only a few specialists in the organization.

Together, enterprise search and classification provide an initial response to the information explosion. Classification complements search by enabling browsing of large repositories based on intuitive categories of information. Like search, classification is quickly becoming an important infrastructure technology.

Once the infrastructure for processing semi-structured documents is in place, there will be an explosion of applications, such as *discovery and social networks* (depicted in Figure 1). These will offer opportunity to extract meaning and business value from our growing repositories of semistructured documents and information.

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.*

© 2003 Robert Blumberg and Shaku Atre