Clear Forest: Cutting through Content Clutter

Taxonomy has become one of the hot concepts in organizations. A taxonomy, according to Inktomi, is “a hierarchical arrangement of topics and subtopics that clearly shows the relationships between each one.” A taxonomy is sometimes called a directory, an informational hierarchy or a topic tree. As Inktomi notes, “A taxonomy is the empty structure, whereas a directory includes the topics and their associated documents.”

With digital information proliferating, tagging and classifying the data is a big job even for small organizations. Once the data have been tagged and classified, users want to know what is in the collection of information. As a result, taxonomy begs data mining and analytics. Yahoo style directories and visualizations show a user what information is available. The realization that digital content requires more than typing a phrase such as “automobile safety” and working one’s way through long lists of Web sites.

In the heady days of the 18th-century, human indexing and filing solved many retrieval problems quite elegantly. But when the volume of information to be indexed grows by terabytes every six months, legions of indexers are required. Individual Web surfers cannot slog through more than one or two pages of documents before fatigue sets in. Even the oligopolistic giants of commercial publishing have to look for a way to make sense out of the information in their systems.

Within the last two years, there has been a growing interest in software that can—mostly without much human assistance—examine a document, link it to one or more appropriate classifications, generate metatags for each document, and provide ways for a user to discover what information resides in a digital collection. A document can be a Word file, a PowerPoint presentation, electronic mail, data in an Oracle database, or any other type of digital object found on individual computers, in small offices, or across a global enterprise.

Reed Elsevier Group recently licensed Clear Forest’s software to assign, update, retag, and perform other feats of indexing legerdemain at blindingly fast speed. Consider the scale required to process 11 million Medline citations. With Clear Forest’s technology, Reed Elsevier’s managers can process digital content to meet the needs of specific customers or create new collections of information quickly. Because market opportunities may be short lived in today’s volatile economic climate, speed is vital.

According to Barry Graubart, one of Manhattan-based Clear Forest’s senior executives, “Elsevier will be creating new content products. But Elsevier will use Clear Forest’s blend of search and analytic tools to allow new relationships within data to be uncovered and those links displayed graphically.

According to Reed Elsevier, Clear Forest technology will be used for ScienceDirect (www.sciencedirect.com), MD Consult (www.mdconsult), as well as other electronic properties that will be rolled out in 2003 and beyond.

Says Mr. Graubart, “Reed Elsevier publishes more than 1,800 journals and 2,200 new books per year. Clear Forest provides a flexible way to process, locate, and manipulate information. Clear Forest’s combination of technology and easy-to-use tools allow editors and customers to extract deeper specialty knowledge from existing content.”

Clear Forest, founded in 1998, is not alone in the specialized discipline of taxonomy or what might be called “next-generation search technology.” Factiva makes use of Wordmap (Bath) and Mohomine (San Diego, California) technology to classify, sort, and tag business information and news for its customers. Verity, Inc. (Sunnyvale, California) includes a taxonomy toolset. Stratify, Inc. (Mountain View, California) offers taxonomy tools that combine automated processes with the work of human indexers.

Companies providing taxonomy software focus on the cost of human indexers. However, the major benefits pay continuing dividends. Mr. Graubart adds, “Taxonomy software provides cost savings. Over time, our customers tell us that the consistency of the tagging is of significant value. Time is money, and when hundreds of users are saving a few minutes to several hours each day, the value to the company is evident.” ClearForest’s technology allows their customers to apply objective, well-defined criteria for tagging documents. Mr. Graubart notes, “The obvious advantage is uniform tagging policy across all documents. The variability—some may say haphazardness and subjectivity of manual tagging—is reduced. Because Clear Forest uses its proprietary semantic approach, a Clear Forest system can automatically adds new entities such as specific company's or executive's names to its dictionary, eliminating the need for the company to make the updates manually.”

Clear Forest’s technology is a closely-held secret. “Clear Forest is not one algorithm,” says Mr. Graubart. “We have engineered solutions that address the difficult problems associated with classifying, enhancing, and tagging digitized information. For example, Clear Forest can extract information of different granularity depending on the user’s requirements. We have a conceptual control panel that can extract useful information ranging from gene pathways to scientist - organization affiliations as well as related documents. Clear Forest’s technology uses what the company calls *hybrid tagging*. The approach is to apply a combination of semantic, statistical, and structural analysis to content. “Content processed by our technology is tagged, but the value comes from the discover of new relationships among pertinent entities, facts, events and relationships. These linkages are often buried deep within text or scattered across different types of documents.

One of Clear Forest’s most interesting capabilities is processing the types of digital information that are found in a typical office. ClearForest has developed a proprietary Intelligent Hybrid Tagging technique that reads content from virtually any source.

Unlike other tagging tools, which only address structured data that have been converted to Extensible Markup Language, ClearForest extracts intelligence from unstructured data. Clear Forest can process hundreds of thousands of e-mail messages, documents, and news feeds from multiple locations.These materials are tagged and made available to the Clear Forest user.

The developers of Clear Forest’s technology are Dr. Ronen Feldman and Dr. Yonatan Aumann. Dr. Ronen is a data-mining pioneer, who has been labeled the “father” of text mining by some colleague. When not in Clear Forest’s office, he is a senior lecturer in the Mathematics and Computer Science Department of Bar-Ilan University in Israel.

Dr. Yonatan Aumann, a co-founder of Clear Forest, is an expert in distributed computing and algorithm design. He oversees strategic product development for ClearForest and holds a senior lecturer position at Bar-Ilan University, with expertise in algorithm design and analysis.

Clear Forest’s offices in Manhattan showcase the award from the KDD Cup 2002. Each year a

data mining competition held in conjunction with the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (See http://www.biostat.wisc.edu/~craven/kddcup/) “In head-to-head competition among the leading players in taxonomy and knowledge discovery. We demonstrated that other systems cannot get at the same information we can. Our secret ingredient is semantic-linguistics analysis. We leverage our algorithms to extract facts, events, relationships. The KDD Cup suggests that we are the best.”

Mr. Graubart continues, “In 2002 the competition included two tasks that involved data mining in molecular biology domains. The first task focused on constructing models that can assist genome annotators by automatically extracting information from scientific articles. The second task focused on learning models that characterize the behavior of individual genes in a hidden experimental setting. Clear Forest performed well across this suite of tests.”

“Reed Elsevier representatives followed up immediately. Their need was to have a complete Clear Forest implementation that they could maintain, operate, and extend as required. Our ClearLab system provides that type of control. When Reed Elsevier management learned that we had a stable system that gave them the environment to create their own intellectual property on the Clear Forest technology, the procurement team realized they could differentiate their product portfolio and build a barrier to competition.”

Clear Forest is one of a handful of new products that straddle traditional search and retrieval technology, indexing and classification, and knowledge discovery. Other companies with a complete solution include the Xerox PARC spin off Inxight Software Inc. (Sunnyvale, California) and Convera Corporation (Vienna, Virginia). Convera, formerly Excalibur Technologies, has strengthened its executive team by adding Claude Vogel, the founder of Semio’s taxonomy system, as its Chief Scientist.

The Clear Forest team has a number of new products moving to launch, including a new version of Clear Research, the company’s core tagging product. “Customers want easier integration with other enterprise software. We are including features to allow customers to use Clear Forest in environments where structured and unstructured data are needed,” says Mr. Graubart.

Clear Forest is not a search software provider. “Search is a very limiting concept,” notes Mr. Graubart. “Clear Forest is not a search company. Our technology allows better search and better insights through research and content navigation. Clear Forest provides users with a way to connect the dots among items of information. The need to move beyond simple search is evident in a wide range of organizations, from the intelligence community to chemical companies, from consulting companies to traditional publishing operations. Our technology helps make search better and supports knowledge discovery and data mining with sophisticated analytics.”

Clear Forest has grown to 60 employees. The company has offices in three locations. The headquarters is in New York City, research and development in Israel, and a Federal sales office in Washington DC. The “taxonomy space” is becoming increasingly competitive. Promising approaches from the once-promising Datops SA (Paris, France) are reminders of the realities of today’s marketplace.

With information growth continuing at mind-boggling rates, Clear Forest is among the business intelligence leaders of tomorrow.