The Early Iterations of Document Mining

A promise to make enterprise knowledge accessible

By Ralph H. Sprague, Jr., Ph.D.

Only 12 percent of organizational knowledge reside in databases as compared to 46 percent in documents that cannot be mined as easily.

Ralph H. Sprague, Jr., Ph.D.
Professor
Decision Sciences
University of Hawaii

During this decade, vast stores of data from point of sale and other business-critical areas have been accessible to business strategists for sophisticated analysis using powerful data mining tools. By successfully comprehending large quantities of data records, these systems permit executives to discover previously unknown relationships and patterns that can provide a competitive advantage.

Yet, only 12 percent of organizational knowledge reside in databases, according to the Delphi Group, a Boston-based market research firm, as compared to 46 percent in documents that cannot be mined as easily. The reason: databases structure data into categories or cells that facilitate queries and other forms of knowledge extraction. Documents are unstructured and offer no comparable inherent means for mining their contents.

Since information work is becoming the true engine of economic wealth, the demand to apply "mining" technology to analyze complex information in documents has been building considerably in recent years. Information overload is an increasing problem; it now promises to overwhelm organizations that do not confront it, while strongly benefiting those that mobilize technology to manage it.

To meet this pressing need, a number of document mining systems have recently reached the market that facilitate these efforts, mainly for text mining. Meanwhile, many leading researchers are seeking to improve upon current techniques to address what we call "semantically rich" documents—multi-media documents containing text, graphics, photographs, audio, video, formatting, structure, and/or hyperlinks.

Already in research labs, scientists have devised systems for recognizing spoken words in audio documents, finding video segments that match pictures and clustering documents based upon the images they contain. Sophisticated scanning software can recognize the format and placement of diagrams, text blocks, photographs, and other visual features in a document.

Eventually, researchers believe that document mining technologies will address every component of semantically rich documents to deliver, for documents, what data mining tools today provide for databases.

Making Unstructured Documents Behave Like Structured Data

Document mining is the process of analyzing a document or set of documents to understand the contents and meanings they contain. A critical challenge in developing these technologies is in overcoming the inherent lack of structure in documents.

Part of the approach to document mining is to provide structure that permits unstructured documents to be queried much like structured data. The user seeks information based upon a concept or idea—the "entity" of interest. The attributes, the equivalent of a data record, may be sets of words or images which describe the idea. All the attributes needed to describe the entity might be a logical paragraph, and so on. Unfortunately, due to the ambiguity of language, these structures do not hold up as effectively as they do in databases.

Consequently, to pick up where structuring efforts leave off, a range of technical capabilities are applied to the task of document mining. These include:

- automatic language identification
- enhanced search and retrieval
- identification and extraction of metadata to form a data record or annotation to represent the document by identifying key features, such as proper names (people, places, and organizations), multiword terms, abbreviations, and currency amounts
- automatic summarization of document content
- automatic categorization according to categories established manually or by defining a sample document which represents the category
- clustering, a fully automatic process which groups documents based on their contents and labels them according to key terms shared by the clustered documents. This provides an overview of document collection contents, identifies hidden similarities among documents, and speeds the process of finding similar or related information in a document collection
- genre identification separates different types of

documents based on the characteristics of the language, format, and content. This enables, for example, distinguishing between news articles and research reports on the same topic

- visualization of concepts and relationships in a document and between documents for faster and easier comprehension

The patterns by which these varied technical capabilities are applied bear much in common with the process steps involved in data mining. The first step is to capture or extract the data and clean or "scrub" it by rationalizing coding conventions. The document mining corollary includes tagging target words and phrases; stemming, or recognizing the root words so as not to be fooled by tricky constructions; and creation of a metadata index of the document's content in a separate, searchable file.

In data mining, the second step is aggregation of the pertinent data, summarization of the aggregated results, and visualization of the data and their relationships through graphs, charts, and other means. Similarly, document mining's second step includes summarization of content, visualization of relationships among documents via graphically oriented user interfaces, clustering of like documents, and categorization of document topics.

The third step of data mining makes use of sophisticated statistical techniques such as regression, correlation, and cross-tabulation to analyze the contents and relationships in the database. In document mining, this third step of intensive analysis of the content of documents and document collections is just emerging.

Document Mining Applications

Most document mining efforts today focus exclusively on text and often are referred to as "text mining" or "text data mining." A popular application for some document mining capabilities is in making email more efficient. Categorization, for instance, can automatically recommend folders to which incoming email should be filed. Clustering can analyze incoming email to discover common themes that may otherwise be difficult to discern. Specialized search functions might be able to isolate required actions or deadlines outlined in incoming email.

Applying these same tools, many enterprises are able to better manage their strategic assets. For instance, a European electric utility analyzes customer feedback, market data, media coverage, customer comments, and market strategy reports as a whole data set, as part of ongoing efforts to review and establish customer policies and procedures. The approach gives the utility a holistic view that would otherwise be difficult to achieve.

Similarly, Eaton Corporation, which markets engines, transmissions, and thousands of other components in vehicles and power plants, uses document mining to automatically monitor technology shifts, emerging competitors, and governmental regulations. Thus document mining provides the company with broad business intelligence.

Other enterprises apply these tools to managing specific divisions or tasks. One medium-size engineering firm uses document mining to analyze internal research and lab reports as a means of exploiting previous successes and avoiding previously experienced pitfalls and blind alleys.

A related application is management of intellectual property. Document mining techniques can help analyze vast patent repositories, highlighting clusters of similar patents, or ranking patents according to areas of interest. Such analyses support research and development (R&D) planning, competitive intelligence, licensing management, and strategic market research.

In the legal profession, document mining is successfully being applied to the legal process of discovery, which requires an examination of large quantities of documents to find occurrences of specific topics, concepts, and words that will be relevant in a trial.

Clearly, many enterprise activities can benefit from this capability to select relevant documents, cluster them into relevant topics, and present the relationship among them in an easy-to-grasp visualization.

The Future of Document Mining

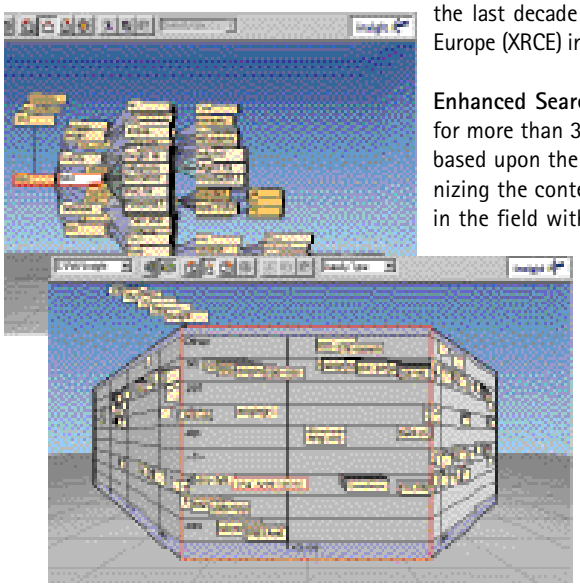
Semantically rich information contained in documents will become increasingly important as the information age progresses. The phenomenal growth of the Internet and intranet guarantee it. Strong, sophisticated technology will be needed to manage, analyze, and understand these information resources.

A plethora of products now are on the market to handle parts of this process. Most will not survive. Simple file servers, office suites, browsers, search engines, and Web crawlers soon will become commodities. The value added in the next generation of products will be the integrated technologies that change how people master the information in their document collections.

The value added in the next generation of products will be the integrated technologies that change how people master information in document collections.

Xerox Provides Leadership in Document Mining

A great deal of the current technology for document mining is based upon the work of The Document Company Xerox researchers over the last decade at the Xerox Palo Alto Research Center (PARC) and the Xerox Research Centre Europe (XRCE) in Grenoble, France. Following are examples of Xerox technology currently available.



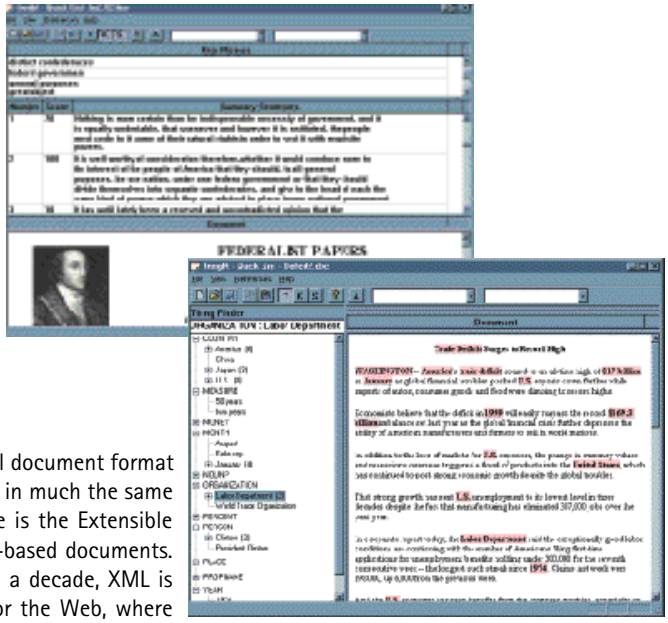
Enhanced Search and Retrieval. Search and retrieval based on keywords has been prevalent for more than 30 years. More recently introduced technology using linguistic analysis, which is based upon the structure of language, improves the process greatly by more accurately recognizing the context within which words are used. Inxight, a Xerox spin-off company, is a leader in the field with its LinguistX Platform (LXP). Most of the popular search engines on the Web use LXP as the pre-processor for their search engines.

Language Identification. This capability, available as part of the LXP platform, automatically recognizes foreign languages and estimates the percentage of each language in a multilingual document. Because LXP's underlying linguistic analysis is based on the structure of languages, the system supports the full range of document mining functions in several languages.

Summarization. This technology plays a major role in overcoming the information overload problem and also is based upon linguistic analysis, in which Inxight has expertise. The firm markets two products, Summarizer and Summarizer Plus, which provide "indicative summaries"—abstracts that indicate content.

Content Extraction. A new product by Inxight, Thing Finder, extracts and identifies attributes, topics, and key phrases in a document.

Visualization. Xerox PARC has a long history of leadership in visualization technologies, having developed the metaphor that is the basis of today's Microsoft Windows and Apple Macintosh operating systems. This research has been ongoing, and Inxight has applied it to a range of visualization products. They include Hyberbolic Tree, Perspective Wall, Cone Tree, and Table Lens, which are embedded in Xerox products like Visual recall, as well as a variety of third-party offerings.



In addition, categorization, clustering, and genre identification technologies are under development within the Xerox family at PARC, Inxight, and the Grenoble laboratories.

Will Future Document Formats Include Structure?

One way to ease the document mining challenge is to use a digital document format that incorporates structure, permitting it to be queried and mined in much the same way as databases. Perhaps the leading candidate to fill that role is the Extensible Markup Language (XML), a specification for tagging text in Web-based documents. While other tagging methods have been available for more than a decade, XML is gaining widespread interest because it is designed specifically for the Web, where structured documents are expected to have significant value.

XML is derived from the same Standardized Generalized Markup Language (SGML) that underlies the current Web-standard format, HTML or Hypertext Markup Language, but is not a replacement for HTML. Where HTML defines how a document is displayed, XML interprets the text. For example, an English literature document might tag Thomas Wolfe as author, facilitating searches by eliminating responses that list others with the same name.

To date, XML usage is limited due to a lack of browser support, the relative immaturity of supporting development tools, and the need to develop tagging standards. Clearly, however, the technology targets an area with tremendous potential. ■