BUSINESS APPLICATIONS OF UNSTRUCTURED

Crunching unstructured text rather than numbers is the basis for a new class of high-value applications ranging from regulatory compliance monitoring to business intelligence.

A widely touted IT factoid states that 80% of the information produced by and contained in most organizations is stored in the form of unstructured data. Most of it is text (such as memoranda, internal documents, email, organizational Web pages, and comments from customers and from internal service personnel), and most of the applications that reflect the value of unstructured data are able to process it. Although unstructured data takes other forms, including images and audio, here I focus on the applications, technologies, and architectures for unstructured text acquisition and analysis (UTAA).

Prior to the legal necessity of monitoring regulatory compliance, the data was largely ignored by both corporate IS departments and corporate

TEXT

By William L. Kuechler

management for several reasons. First, its value was considered dubious. Second, even if it had value, the expertise required to tap it (such as latent semantic analysis and other statistical forms of corpora analysis) and natural language processing (NLP) were outside the domain of mainstream IS. Third, the terabytes of data storage and megaflops of computing power required to implement large-scale unstructured data analysis applications were, until the advent of blade processing and terabyte disk arrays, prohibitively expensive. Here, I provide an overview of a rapidly maturing area of technology poised to be an integral part of most organizational IT portfolios.

The storage and processing power required for UTAA applications are today relatively inexpensive. In addition, three powerful business drivers are pushing unstructured data acquisition and analysis applications to the foreground of IT management. The most compelling, for the organizations to which they might be applicable, are the legislative mandates that help monitor organizational behavior and communication, including the Sarbanes-Oxley Act of 2002, SEC Rule 17a-4, and the Health Insurance Portability and Accountability Act of 1996. In all of them, IT carries the primary responsibility for ensuring that private information does not become public. The maintenance of personal privacy requires attention to access-security issues, as well as ongoing monitoring and analysis to ensure that even authorized users behave

Less urgent but more pervasive than government regulation as an influence in UTAA applications is the almost complete assimilation of sophisticated Internet search into corporate life at all levels.

responsibly and ethically. Larger organizations have begun to capture and monitor all information flowing over or out of any portion of the organizational intranet, including chat, email, attachments, and Web pages. In the event of an investigation, experts advise that documents be stored in a system that allows forensic search, since evidence of due diligence is required to forestall legal action against the organization's employees and against the organization itself. In practice, these mandates cannot be satisfied without computerized UTAA applications.

UTAA Application	Textual Data Source
Business intelligence	Web, industry blogs, online databases
Customer relationship management	Customer feedback, help desk reports
Regulatory compliance	All internally generated electronic documents
Intellectual property management	Web, copyright and patent databases
Call support (help desk applications)	Call documentation, customer feedback, email, online manuals
Accounts payable/receivable analysis	Invoices, customer and vendor correspondence (used frequently with traditional structured data mining and analysis)
Legal department support	Legal databases, specific streams of organizational communications (such as customer communication, internal email)

Table 1. Common UTAA applications and their data sources.

Less urgent but more pervasive than government regulation as an influence in UTAA applications is the almost complete assimilation of sophisticated Internet search into corporate life at all levels. Familiarity with sophisticated search techniques for a huge corpus of Internet data (notably Google) has increased demand within organizations for the same level of access to organizational data. Gartner Research has determined that "Once Internet search became available, people expected that same level of availability in their business lives" [5].

Finally, an increasing number of applications that are heavily dependent on UTAA (such as Web scanning for business intelligence, customer relationship management, and service-desk assistance) have moved from unproven, high-risk pilot projects to valued systems in many corporate IS portfolios. Many CIOs are increasingly receptive to strategic UTAA applications suggested by internal operations staff or available from specialty software vendors. Table 1 lists some commercial off-the-shelf (COTS) applications based on an analysis of unstructured or partially structured textual information. For example, computer analysis of customer and dealership comments and email messages has become a common technique used by automobile manufacturers to anticipate potentially widespread mechanical problems that may require recall notices [11]. And analyzing feedback from customers, together with the ability to mine the Web, including online product reviews and industry-specific blogs, is the cornerstone of today's business intelligence applications.

Architecture for UTAA Applications

While the scope of UTAA applications ranges from patent search to predicting next year's cuff style for men's pants, the applications share a common structure (see Figure 1). The functional architecture outlined in the figure synthesizes multiple, usually partial, architectures found in vendor documentation and research literature. Few if any organizations require all the functionality in all the modules in Figure 1, though it serves as a road map to UTAA and is general enough to trace the operation of an individual application or as an enterprise architecture for an organization's UTAA portfolio. The architecture is logically divided into three subsystems: document storage and administration, labeled (1) "enterprise document management" (EDM); primary processing, labeled (2) "application analytic processing" (AAP); and support, labeled (3) "application context administration" responsible for linguistic and domain-descriptive data used in text analysis.

Enterprise document management.[1] The primary EDM functions are long-term storage of source text

[1]The term "enterprise document management" indicates enterprisewide document storage administration and optional format consolidation; "enterprise content management" is an alternative term common in the content-management-vendor literature, as well as in practitioner and academic literature, defined in such a variety of ways as to be confusing [3].

documents and administration and retrieval of the documents. Long-term storage is required for regulatory applications and for the computationally intensive processing of documents. Source documents must be specified explicitly, since, as an optional subsystem function, documents may be converted from their source form to a common format prior to being processed. However, the legal status of converted representations is ambiguous, and some formatting information is lost in conversion, so source documents are usually retained, even when conversion is used.

Many vendors supply software specifically for the EDM subsystem; however, there is no industry standard for the functions provided or even for the generic name of such software.[2] Most EDM software is complex, expensive, and intended for the enterprisewide content management required for regulatory-compliance applications. Most regulations do not distinguish among specific forms of communication but rather specify the monitoring and storage of all organizational communication that may contain regulated content.

Application analytic processing. The AAP subsystem is divided into two parallel streams. Following [10], this architecture distinguishes document-based processing from concept-based processing. Simple keyword searches are one type of document-based processing; the intent is for the application to return links to all documents in a corpus matching user-entered keywords, in the same way the Web is searched through Google and Yahoo. Concept-based processing usually analyzes a corpus for relationships among concepts in document content. Links to documents are returned only as corroborating evidence of the concept-based results.

An illustrative concept-based application [2] has created and maintains a database of more than 100,000 expert witnesses for a variety of court cases. Public jury verdict and settlement documents, along with professional license and other Web-available records, were mined to produce the database. This is an obvious example of a concept-based application, since the output is a collection of individual

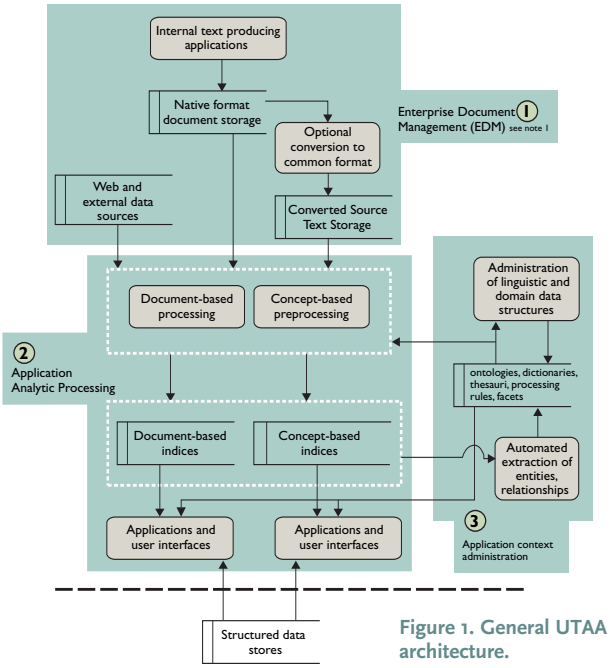

Figure 1. General UTAA architecture.

experts or concepts referenced across multiple documents, not documents or references to the input corpus itself.

Note that either type of application may reference traditional structured databases as part of its processing. This is done in advanced applications either to combine traditional data-mining results derived from the structured data with UTAA results or to draw data from the structured databases for use analyzing textual data.

Application context administration. Many of the algorithms and processing techniques described in the following paragraphs depend on linguistic and domain-specific context information generated, reviewed, and administered by subsystem (3) in Figure 1. A dictionary of "noise words" (such as "a" and "an") assists in the removal of these words from a corpus, as required by many NLP and statistical techniques. Thesauri, together with rule sets, are often needed to disambiguate superficially similar expressions in different contexts.

Document- and concept-based processing often requires a domain ontology. UTAA ontology descriptions are similar to traditional biology taxonomies or to class, superclass, and subclass relationship descriptions in object-oriented programming. Many vendors supply general-purpose UTAA software customized to a particular domain (such as to analyze sales reports and search pharmaceutical journals for undiscovered drug interactions) through the addition of precomposed ontologies, dictionaries, and thesauri.

NLP modules often require a significant amount

[2]DOD 5015.2 is a standard for record management that specifies minimum management, security, and search capabilities for record-management applications purchased for the military. Though exceeded by most commercial packages, it guarantees and defines a base point. The full specification is at www.dtic.mil/whs/directives/corres/pdf/50152std_061902/p50152s.pdf.

of reprocessing and/or customization when applied to text corpora on which they are not already trained. The quality of a UTAA analysis can be altered dramatically through small changes in contextual data (such as ontologies and dictionaries) or in processing rules. Moreover, this data is time-sensitive for many applications (such as when analyzing business trends). The data must therefore be edited by domain experts, usually with help from specialists in linguistic-analysis techniques. Since this maintenance effort is so specialized, in addition to being time-consuming and sometimes tedious, most third-party UTAA application suppliers provide consulting services to help fine tune these domain- and environment-specific data stores.

Techniques and Algorithms for UTAA

For a fuller understanding of UTAA it is useful to look inside the black boxes of the application analytic processing subsystem (2) in Figure 1. Since computational methods for analyzing unstructured text represent a large and growing field [6], I have limited my discussion here to the most basic processing strategies and algorithms:

Document-based processing. Many document-based UTAA applications (such as business-intelligence environment scanning and certain types of intellectual-property management) are retrieval functions, sometimes called "document-based predictive analysis." These applications are basically categorization functions; responding to an ad hoc or prespecified query, they search a document corpus, often with the aid of precompiled indexes to retrieve pointers to documents that are "about," or have been categorized as pertaining to, the query as a whole or to the individual terms in it.

The most significant difference between searching a corpus of unstructured text and a relational database inquiry is that, whereas a relational model demands keys and firmly defined columns that unequivocally determine what the data records are about, the "aboutness" of a text document is fuzzy, at best. A single document can legitimately be about many different things, depending on the context of the search. For example, a recipe for veal scaloppini published in the Web version of a popular U.S. magazine can be about Italian meat dishes in a casual scan by someone looking for a main course for a dinner party or about high-protein diets in a search for American eating habits performed by a medical researcher. It should now be apparent why an entire subsystem of the general UTAA architecture in Figure 1 (3, application context administration) is devoted to maintaining the data structures that provide context.

Figure 2 outlines the general course of most categorization processing. Variations on the basic techniques described earlier are endless, including "learning" interfaces and tag-based indexing, which inserts XML tags for elements into the documents in the corpus. Tag-based indexing can be viewed as a form of preprocessing of the corpus during which portions of the document are mapped (classified) to the XML tags. This mapping can significantly reduce the time required for ad hoc searches and index construction compared to similar operations on an untagged corpus. Unlike the individual-document emphasis of document-based processing, concept-based processing emphasizes analysis across a corpus.

Concept-based processing. Virtually all concept-based algorithms make extensive use of NLP techniques to determine the concepts and relationships in document text. Text can be understood at many levels, as outlined in Table 2. Analyses of text can be made at a single level (such as morphological), but many analyses proceed hierarchically, with the analysis at a lower level of language (such as morphological) providing data for inferences at a higher level (such as syntactic).

Level	Analyzes	Output
Morphological	Words and allowed variants	Terms in documents
Pragmatic	Contextual meaning, frequently implicit	Intention of textual units
Semantic	Relationship of words to world knowledge	Concepts and relationships
Statistical	Co-occurrence of terms	Strength of term relationships
Syntactic	Word ordering	Structural relationships among terms

Table 2. Levels of NLP analysis of textual material.

Latent semantic analysis (LSA) is a useful concept-discovery technique used in many variations in multiple UTAA applications. Originally developed for the conceptual indexing of documents [4], LSA and variants are now used for purposes as diverse as Google's category-discovery algorithm and automated scoring of student essay examinations. Unlike most document-based categorization techniques, LSA does not use a preexisting concept set. LSA discovers concepts from an analysis of the document set of interest by processing every nontrivial word in a corpus into huge occurrence matrices. The rows in these matrices or vectors can be thought of as

Besides advising against consolidated data storage, I suggest evaluating potential UTAA applications through the same heuristics used for cost-benefit evaluations of other types of business systems.

expressing what the document terms are about. LSA's extraction of meaning is described in [4], which says "...LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains."

Other common concept-based analysis techniques are statistical in nature, trying to infer the significant concepts in a document or corpus (in simple terms) and counting and weighing word or phrase occurrences or co-occurrences. As in document-based techniques, the field of statistical analysis is advancing rapidly, and many recent algorithms have been patented.

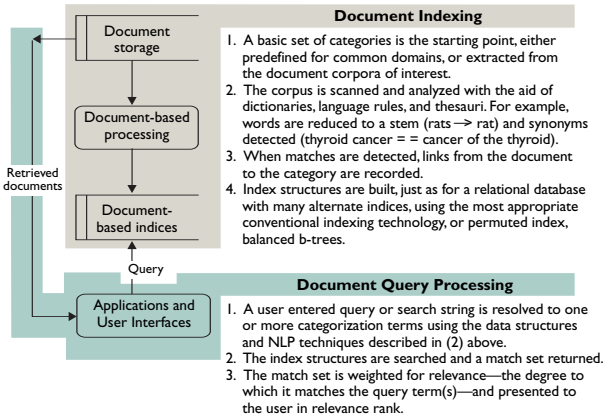

Figure 2. Document-based indexing and query processing.

Make, Buy, Implement

Although they process nontraditional data types with possibly unfamiliar algorithms, UTAA applications are basically business software applications. Thus, much of the conventional wisdom pertaining to software make-buy decisions and vendor selection applies. Here, I outline (based on the empirical observations of early adopters) where the conventional wisdom breaks down and attempt to fill the breaks with some UTAA-specific guidelines.

It is useful to position UTAA within the traditional MIS taxonomy of applications. As far as the increasing number of UTAA application vendors is concerned, UTAA is synonymous with knowledge management; indeed, the trade publications and trade shows that describe themselves as serving the knowledge management industry deal with little more than UTAA systems. However, depending on the class of application, UTAA applications are best viewed as fitting into one of two more-traditional categories in an organizational IS portfolio. If the application is directed toward regulatory compliance, it should be considered a computing infrastructure component, much like email or word processing, because all current regulations require organizationwide monitoring and auditing to ensure compliance. All other UTAA applications tend to be domain specific and thus better categorized as strategic departmental applications.

Classification into departmental or infrastructure applications is also helpful in determining whether or not some form of EDM is required or even desirable for a particular application (subsystem 1 in Figure 1). It is with respect to EDM that early field observations of UTAA applications most strongly veer from the conventional wisdom on structured-data IS. Business computing over the past 50 years has taught and confirmed that, as far as structured data-based systems are concerned, information silos are nonproductive. In contrast, practice to date with UTAA systems, other than those in regulatory compliance, has shown the opposite—that attempting to consolidate multiple unstructured data sources for multiple, domain-specific UTAA applications is expensive, nearly impossible to administer, and interferes with, rather than enhances, functionality [3, 8].

Two aspects of unstructured data contribute to application isolation: unstructured data comes in an array of formats, and unstructured data is topically more diverse than structured data. While it may make sense to integrate quantitative (numeric) information from multiple organizational sources, there is no immediately obvious reason for forcing sales memos and purchasing correspondence (both textual data) into the same format and data store. The dis-

Pushed by legislation and pulled by enhanced technology and heightened awareness of business benefit, UTAA thus seems poised for rapid entry into the mainstream of business information systems applications.

parate nature of unstructured data makes it difficult to realize economies of scale from consolidation. Moreover, distinct application operations (and thus different training programs) and modes of search are actually a benefit to the radically different functions served by UTAA.

Based solely on preliminary field reports from early adopters, a number of consultants now recommend against naive (monolithic) consolidation [3, 8]. However, an alternative concept called "enterprise content integration" (ECI) shows promise [8, 9]. As an application of general information-integration principles [7] to the text domain, ECI represents a federated approach to unstructured data integration that allows each UTAA application to remain autonomous while using a middleware "adapter" to connect it to a centralized metadata administration application (see Figure 3).

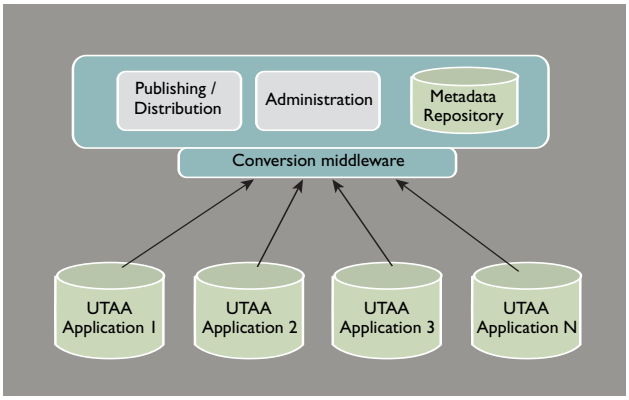

Figure 3. General ECI (federated) architecture.

The same criteria apply to the make vs. buy decision for UTAA as in other software applications; however, due to the relatively short history of the field, information on applications and construction of applications is difficult to find. Here are three widely cited resources:

- The KMWorld Buyers guide for an introduction to COTS applications (www.kmworld.com);
- The Stanford NLP Web site, which lists open-source software libraries of UTAA software subroutines for use in in-house development (nlp.stanford.edu/links/statnlp.html); and
- The General Architecture for Text Engineering project at the University of Sheffield (gate.ac.uk).

Besides advising against consolidated data storage, I suggest evaluating potential UTAA applications through the same heuristics used for cost-benefit evaluations of other types of business systems. Where regulatory compliance is at issue, some form of ECI may be warranted; however, as this application is new, standards are relatively immature. UTAA technology is leading edge and thus higher risk than analogous applications (such as data mining and multidimensional analysis) performed on relational databases.

Conclusion

This discussion represents a sound basis for appreciating the technical, environmental, and business trends that affect the future of UTAA in the business environment. One technical trend toward greater use of UTAA applications involves dedicated hardware for search and retrieval. Google is one of the first vendors to package the hardware and proprietary software required for a complete intranet search engine, ready to be hooked up to a network to crawl and index pages and respond to search requests; if the market responds, multiple vendors can be expected to follow.

Another technical push will follow from the increasing sophistication of NLP techniques. Research on unstructured text analysis is well funded for the foreseeable future due to increasing commercial exploitation of the Web and the security benefits that follow from being able to scan and interpret vast quantities of electronic communications. The direction most researchers feel will yield the most direct research benefit to UTAA involves increased NLP understanding.

Finally, many experts believe IT managers in the U.S. have seen only the tip of the regulatory iceberg, and the UTAA applications used to monitor compliance are typical of the type of application that will be required in the future. Recognizing the pervasiveness of compliance issues, some organizations have

assigned compliance as an IT staff responsibility, with significant software (much of it UTAA) and infrastructure requirements [1]. Driving the increase in regulation are security (nationally and as a defense against computer fraud) and privacy issues, an area in which the U.S. is significantly behind the European Union; international businesses are being pressured to increase compliance with EU data security standards [12]. Pushed by legislation and pulled by enhanced technology and heightened awareness of business benefit, UTAA thus seems poised for rapid entry into the mainstream of business information systems applications. C

References

1. Britt, P. ECM...no slowdown in sight. *KMWorld 16,* 3 (Mar. 2007), 8–11.
2. Dozier, C., Jackson, P., Guo, X, Chaudhary, M., and Arumainayagam, Y. Creation of an expert witness database through text mining. In *Proceedings of the International Conference on Artificial Intelligence in Law* (Edinburgh, Scotland, June 24–28). ACM Press, New York, 2003, 177–184.
3. Howard, J. ECM: Don't buy it. *CMS Watch* (Mar. 26, 2003); www.cmswatch.com/ECM/.
4. Landauer, T., Foltz, P., and Laham, D. Introduction to latent semantic analysis. *Discourse Processes 25,* 2–3 (1998), 259–284.
5. Moore, A. The universe of search. *KMWorld 13,* 7 (July–Aug. 2004); www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=9564.
6. Munoz, R. and Montoyo, A. Introduction to a special issue on advances in natural language processing. *Data and Knowledge Engineering 61,* 3 (2007), 403–405.
7. Park, J. and Ram, S. Information systems interoperability: What lies beneath? *ACM Transactions on Information Systems 22,* 4 (Oct. 2004), 595–632.
8. Rosenblatt, B. Enterprise content integration: A progress report. *The Seybold Report 3,* 10 (2003); www.seyboldreports.com.
9. Silver, B. Content: The other half of the integration problem. *Intelligent Enterprise* (Oct. 1, 2005); www.intelligententerprise.com/toc/;jsessionid=NRA1DMRLNKQGSQSNDLOSKHSCJUNN2JVN?day=01&month=10&year=2005.
10. Tan, A. Text mining: The state of the art and the challenges. In *Proceedings of the Pacific-Asia Conference Knowledge and Data Discovery 1999, Workshop on Knowledge Discovery from Advanced Databases* (Beijing, Apr. 26–29). Springer, Berlin, 1999, 71–76.
11. Vasilash, G. Diagramming sentences can save your company billions. *Automotive Design and Production* (Aug. 10, 2004); www.autofieldguide.com/articles/.
12. Wu, J. Business intelligence: What you need to know about the EU data privacy directive. *DM Review* (May 26, 2005); www.dmreview.com/article_sub.cfm?articleId=1028642.

William L. Kuechler (kuechler@unr.edu) is an associate professor in the Department of Accounting & Information Systems in the College of Business at the University of Nevada, Reno.