Finding Value in Text Analytics

A white paper prepared for

Text Analytics Summit 2006

Boston, June 22-23, 2006

(www.TextAnalyticsNews.com)

Seth Grimes

Alta Plana Corporation

Table of Contents

Executive Summary

Text analytics software has built on early-adopter, niche success to define a broader role in solving common and essential business problems. But while potential usage is huge, market penetration is as-yet small. The potential will be realized as vendors, researchers, and implementers articulate the breadth and depth of text analytics solutions by further and more simply explaining how text analytics interplays with and enhances established operational and analytical technologies. Essential points can be distilled from case studies that demonstrate what potential users can expect and to guide them in evaluations and planning. And analysis of user experiences will foster the development of best practices to support widespread enterprise implementation.

Text analytics solutions complement and extend Business Intelligence (BI) and data mining. While the solutions build on content management and text search rather than on databases and queries – they work with documents, the most common form of "unstructured" data, rather than fielded, numeric data – they share the broad goals of their analytics cousins. Both forms of analytics detect patterns and relationships in data to create knowledge and enable business-process automation.

Text analytics starts with source-document acquisition. Lexical and linguistic analyses help structure content for feature extraction and statistical data mining that can classify information and detect clusters, linkages, associative and predictive rules, and other forms of relationship. Visualization interfaces, especially network diagrams that display entity and concept interrelationships, as well as browsers and navigators that link extracted information back to source texts, deliver results for interactive exploration. The goals are to *operationalize* both text- and number-based analytics and to support analytics both as a business process and embedded within line-of-business applications.

Technology rarely sells itself, however, and text analytics is no exception. It has to solve business problems. Text analytics can find missed opportunities by tapping previously inaccessible textual data sources and applying techniques that cannot be performed manually. Speed, throughput, and breadth of coverage in terms of volume and languages far exceed those of human operators. And text analytics combined with BI, Web mining, and survey analysis can improve results – provide "lift" – and help automate manual processes, saving time and money.

Case studies describing early-adopter experiences can be mined to help prospective users assess how text analytics could work for their organizations, evaluate technology options, design solutions that integrate the new capabilities and optimize related business processes, estimate expected return on investment (ROI), and understand best practices that can be applied at their own organizations.

Continued text-analytics success seems inevitable given today's hypercompetitive business climate and the commonly cited estimate that 80 percent of organizational information is in textual form. The technology is robust and market acceptance is growing rapidly. Vendors are improving at explaining the technology and its applications. Formalizing the message will take the market to the next level. The communications challenges lie in demonstrating value to potential users and in articulating best practices that will guide them in their own evaluations and implementations.

1 Introduction: Finding Value

Text analytics is technology and process both, a mechanism for knowledge discovery applied to documents, a means of finding value in text. Like other technologies and technology-enabled processes, text analytics creates value by offering capabilities and approaches that improve on existing solutions, solve new problems, and pay for themselves by generating new opportunities and revenues and lowering costs.

This paper is about the business value of text analytics, about seeing the benefits that the technology and processes can bring to adopters. Rather than explain text analytics technology or its market in great depth, it focuses on value and suggests how organizations can assess text analytics' ability to meet business needs and how they should evaluate options and implement the new capabilities. It proposes a more formal approach than most organizations have taken to date.

To find value in text analytics, we do have to understand the fundamental technology and how it is applied. We can then develop a valuation framework. These elements are covered in the sections under Technology, Applications, and Valuation. The User Stories sections profile the market and examine a selection of user experiences. The Development of Best Practices attempts to distill basic lessons from the user experiences evaluated in light of the valuation framework.

The components of this paper – a start on a rigorous but flexible valuation framework, sketches of user stories, and an effort to derive evaluation and implementation best practices – should complement technology-centered papers and case studies to suggest how text analytics can be transformed into a robust, mature analytical discipline.

2 Technology, Applications, and Valuation

2.1 Technology

The term *text analytics* describes a set of linguistic, lexical analysis, pattern recognition, entity extraction, tagging/structuring, relationship analysis, visualization, and predictive techniques. The term also describes processes that apply these techniques, whether independently or in conjunction with query and analysis of fielded, numerical data, to solve business problems. These techniques and processes offer an ability to discover and present knowledge – facts, business rules, and relationships – that had previously been locked in textual documents, impenetrable to automated processing. They enable computer programs to interpret and act on human communication in its native forms.

Text analytics starts with text acquisition from file systems, from content-management, database, e-mail, and other servers, and from networked sources including Web sites and databases and even data streams. Lexical and linguistic analysis, with tagging and entity extraction, help structure document content for further processing using established data mining techniques, accounting for special characteristics of text such as, considering the number of *features* in a typical document set, high dimensionality. Visualization interfaces, especially network diagrams that display entity and concept interrelationships, as well as browsers and navigators that link extracted information back to source texts, deliver results for interactive exploration while integration with operational systems remains more ad hoc.

The field of text analytics is linked to the fields of content and knowledge management, which seek, among other things, to organize textual and other "unstructured" data sources and provide methods and tools for capturing, storing, retrieving, and delivering corporate learning. Text analytics extends and often exploits traditional semantic analysis and information categorization methods: thesauri that organize business-domain terminology, taxonomies that provide subject-domain-specific, hierarchical information classifications and ontologies that assert relationships among classified entities.

Text analytics is closely related to search, which similarly aims to index information for retrieval but provides only rudimentary relationship analysis. Text analytics goes beyond search by exploiting statistical data mining and machine-learning approaches to discern clusters, categories, linkages, and predictive rules that apply to source documents and the entities and concepts they contain.

These technologies all deal, rather than (only) with fielded, numeric data, with documents and the features they contain: words, terms, and simple and complex entities such as names, dates, events, and concepts. Some technologies are also used with other forms of "unstructured" data such as audio, images, and video, but given the volume of text in enterprise settings, there's much to be gained by keeping the focus on text. The overarching goal is to detect relevant patterns and relationships in source data, structure that information for use, and enable both interactive exploration and automated handling of the source materials based on analytical findings.

2.2 Applications*

Applications tackle specific business functions or processes. They may be independent of the area of use or they may be linked to particular business domains. Vertical solutions take domain tailoring to an extreme: they fill multiple needs in their domain(s) of use.

The list of business functions that already apply text analytics is long. It includes:

- Chemical/Drug/Gene/Symptom Discovery

Sources include scientific and clinical literature, treatment records and reports, databases of chemical compounds, pharmaceuticals, and medical symptoms, and so on. The goals include extracting relationships among biomedical and chemical entities, genetic markers (e.g., proteins and genes, base sequences), and symptoms – patterns such as "A inhibits B" and "A activates B" and "A is associated with B." In this application, entity extraction recognizes gene, protein, chemical, symptom, and syndrome names and signatures from biomedical and chemical texts based on a domain dictionaries. There is a need to detect temporal relationships and associations and other forms of pattern.

- Health Care Case Management

Sources include clinical-research databases, patient records, insurance and regulatory filings, and regulations. Goals are to enhance diagnosis and reduce misdiagnosis, ensure adequate treatment, promote quality of service, increase utilization, reduce fraud, and control costs.

- Intelligence and Counterterrorism

Sources include news and investigative reports, communications intercepts, documents, and case files, all in a variety of languages. Targets are organizational associations and networks, behavioral/attack patterns, threat assessment, strategy development, tactical evaluation, and event prediction.

- Law Enforcement

Sources include case files, crime and court reports, legal documents, and geographic and demographic data. Goals include detection of crime patterns (temporal, geospatial, and involving persons and organizations) and support of criminal investigations and prosecutions.

- Securities Fraud Detection

Sources include financial and news reports, corporate filings and documents, trading and other transaction records. The goals include detecting insider trading, reporting irregularities, money laundering and illegal transactions, and pricing anomalies.

... and legal discovery and strategy development; patent examination; recruitment

* The Applications section is excerpted from Alta Plana's 2005 white paper, *The Developing Text Mining Market*.

including resume processing; and survey analysis when there are free-form responses. Many of these uses involve linking information derived from text to numerical data.

Text analytics can play a significant role in many business functions independent of business domain. The list includes notably Customer Relationship Management (CRM), where sources include customer e-mail and letters, call center notes and transcripts, and also, potentially, data maintained in CRM systems, and aims are to identify product and service quality issues, to assist in product design and management, and to route contacts.

Lastly, text analytics enables creation of new functional applications such as Reputation Management, which involves gathering and crunching news reports, Web pages, market analyses, correspondence, and other documents; extracting concepts including "sentiments" and scoring criteria and weights; and running analyses. Without text-analytics tools, Reputation Management would be prohibitively expensive and slow and would possess very limited reach. Similarly, Social Network Analysis tools analyze e-mail and other communications, corporate documents and news reports, and other sources to determine connectedness of individuals and organizations, and best paths for routing contact requests. This form of analysis would be impossible without text mining.

2.3 Valuation Framework

A valuation framework is a decision tool that provides a means of assessing the benefits of available options, whether rooted in technology or not. Key points to incorporate into a text-analytics valuation framework include possibilities for process automation, broadened document intake (format, location, language), adaptation to particular subject matter domains with distinctive lexicons, statistically rooted knowledge discovery, and complementing analysis and knowledge discovery based on numerical data. That text analytics supplements or replaces human efforts is an important factor.

Since text processing has heretofore been a largely manual job, human costs come into play. Optimally, of course, text analytics can relieve people of drudgery and free them for higher-value work. For instance, a text mining solution can scan communications traffic for key phrases and patterns, and raise alerts to human operators only when needed. Or a solution can determine which chemical compounds are the most promising for lab testing in fighting a disease, saving researchers from unnecessary, fruitless laboratory trials.

In assessing the potential benefits of text analytics, an organization could adapt this valuation framework or create one of its own.

2.3.1 Missed Opportunity

A valuation framework for text analytics should start with missed opportunity, in view of the commonly accepted figure that 80 percent of enterprise information is in textual form and recognizing the high cost, slowness, and limited reach possible if the sole means of processing textual information is manual. An organization assessing text analytics should ask,

- Are existing methods – typically reliant on search, structured query, and

manual processes – capable of extracting the greatest business value from textual information sources?

- What textual data sources do we now have that we are not exploiting or that we are underutilizing? What would the costs be, including scanning and reformatting, of tapping new sources?
- What additional sources, for instance news services, are available that could extend our reach? Do they offer bundled text-analysis services, or would we acquire information and do text analytics in-house?
- What are the contents of these sources and how would we use information extracted from them?

2.3.2 Enterprise , Competitive, and Customer Value

- How will text analytics enhance the value of our internal operations or of our products or services?
- Can text analytics save us money by standardizing and automating treatment of text and by reducing error?
- Have our competitors pursued text analytics?

2.3.3 Evaluation and Implementation

- Which vendors have demonstrated solutions or case studies for our industry, subject-matter domain, or language(s)?
- Do our current content management, BI, or data mining vendors offer text analytics solutions, either their own or via a partnership or OEM arrangement?
- Are candidate systems feature-complete? If they rely on OEMed technology, what risks and advantages are created?
- How do we integrate text analytics with existing operations and/or analytics?
- What are the licensing and support costs? What professional services, staff reassignment and training, and occasional or on-going consulting does a text-analytics program entail? What is the "total cost of ownership" over time?

2.3.4 Performance Measurement

- What results do we expect and when? That is, what are our goals and the timeframe in which we expect to meet them, the "time to value"?
- How will we measure progress toward our goals?
- How will we quantify benefits, measure costs, and compute return on investment (ROI)?

Text analytics offers capabilities that automate labor-intensive processes, extend the intake and reach of text handling, and provide very high performance and throughput in text processing. Benefits are offset by evaluation and implementation cost including the cost of reengineering business processes to admit automated text processing. Operating costs are balanced by greater efficiency and productivity and new capabilities.

3 User Stories

3.1 User Segmentation

It is illuminating to segment text-analytics users and prospective users by attributes that include:

- Business domain and business function: for instance, pharmaceutical drug discovery, clinical and research medicine, law enforcement and intelligence, manufacturing and quality assurance, customer relationship management and marketing; focus on efficiency and profitability vs. on expanding capabilities;
- Organizational role: Text analytics is most commonly used by researchers and business analysts and embedded in line-of-business applications;
- Type and source of information to be analyzed, for instance the language(s) and use of natural language or a domain-specific vocabulary; acquisition from the Web, from e-mail, from forms and reports and database- and content-management systems;
- The features – the entities and concepts – of interest and the character of their interrelationships and their relationships to field- and record-structured data in relational databases; and
- Vendor relationships: Some users have relationships with text-analytics vendors forged through use of data-mining and other analytical tools.

This segmentation will in turn suggest users' analytical requirements, such as, variously, supervised (categorization) vs. unsupervised (clustering) learning; their interface needs, e.g., for visualizations that support exploratory analysis or for automated, hands-free document processing; and how they would likely seek to integrate text analytics with operational and analytical systems.

Applying a bit of Bayesian reasoning (text analytics being a relatively new field), early-adopter segmentation and user stories should indicate criteria the broader market will apply in assessing solutions.

3.2 Case Studies: An Information Gap

Case studies published by vendors typically outline particular business problems and their solutions and often describe and quantify benefits in terms of new and enhanced capabilities and reduced operational costs.

Case studies rarely cover deployment and integration costs – software licensing, consulting and professional services, training – or ongoing operational costs. Of course they highlight only successes in a world where an estimated quarter of licensed BI software is shelfware and a probable majority of IT projects fail to fully meet business requirements and user expectations. And some vendors provide little more than quotations from selected customers. It would be great if more and better case studies were available, but given the newness of text analytics, companies are looking to the technology for competitive advantage and are frequently not forthcoming.

A few notable exceptions notwithstanding, materials provided by vendors offer little

of value beyond a reference to prospective users. The information gap can be closed by encouraging greater customer participation in conferences (whether dedicated to text analytics or data mining, to knowledge management, or to a particular industry) and by creating sets of typical stories – call them usage patterns – reflecting the spectrum of user requirements. We want to move toward a market like the relational database management system (RDBMS) market with plentiful opportunities for direct contact among users and prospects, relatively simple segmentations (like the RDBMS world's transaction processing vs. reporting vs. analytics) that allow prospects to focus their technology assessments quickly, and standard evaluation and deployment approaches.

3.3 Segments and Styles

I spoke to a sampling of experienced text-analytics users to learn how they had assessed the technology, evaluated options, and developed solutions for their organizations. I discovered that the most useful user division when studying these technology-introduction processes is between, first, users who extended existing data mining and analytical applications to cover text or whose starting point was a search application and, second, users who are building a pure-play text analytics solution from the ground up.

Extending established analytical methods is surely the easiest route to text-analytics. Users who follow this route are already familiar with the statistical concepts involved in knowledge discovery. They know data mining and machine learning and output visualization and understand the importance of auxiliary processes such as data preparation that consume the majority of the data miner's time. Often, as in survey analysis where there are free-form response fields and in call-center-response and warranty-claims analyses where much of the information is captured in "semi-structured" form, moving to analyzing the textual data is a relatively small step. These users appear most concerned with expanding capabilities, doing what they couldn't (affordably) do before, and not overly focused on financial return.

Some organizations seek to add analytics to existing text-centered solutions, extending search and information-retrieval applications. These users are often already familiar with thesauri and taxonomies and with lexical analysis of source texts. Their goals often involve adding new end-user services.

Prospective users jumping fresh into text analytics appear to tend to apply the technology on a much larger scale in projects that tackle new business problems rather than ones that seek incremental improvement. They are also among the least communicative of text analytics users about their experiences, whether because of the secretive nature of their applications (such as intelligence and counterterrorism) or their perceptions of the competitive advantage afforded by text analytics. This reticence makes published studies such as one describing IBM's TAKMI (Text Analysis and Knowledge MIning) system* all the more valuable. These users are among the most likely to have conducted formal vendor surveys and evaluation processes, often with vendor assistance.

We acknowledge the diversity of user backgrounds and styles in providing principles for creation of text-analytics best practices.

* http://www.research.ibm.com/journal/sj/404/nasukawa.html

4 Best Practices

4.1 The Rationale for Best Practices

"Best practices" are generalized principles, techniques, and methodologies derived from theory, academic and industrial research, direct experience, and case studies that guide those who apply them in assessing requirements, evaluating options, and devising strategies for the implementation of new technologies. The application of best practices recognizes that while each enterprise has unique circumstances and requirements, there are very broad areas of commonality among organizations with similar purposes and needs that we should seek to exploit.

4.2 Adapting and Applying a Valuation Framework

Prospective users need to assess how text analytics could work for their organizations, to evaluate technology options, to design solutions that integrate the new capabilities and rework and extend existing business processes, and to be able to compute and report ROI. Best practices call for creation of a valuation framework for assessment of business needs and the potential of the technology for meeting those needs. Such a framework is distinct from and preliminary to creating evaluation criteria.

Assessment starts with examination of existing organizational work practices and business processes. Because technology for automated analysis of text has been late to emerge, best practices should focus on transformation of manual processes rather than on replacement of legacy systems. Best practices suggest managed processes that lead to measurable outcomes that are well aligned with business goals and emphasize integration with established operational and analytical systems.

4.3 The Evaluation Process

The best practices next step, prior to formal product evaluations, is a survey of technical options and of case studies to understand experiences at comparable organizations and to establish evaluation criteria.

It appears that the majority of users who have embraced text analytics may have done only informal, ad-hoc evaluations, responding to individual rather than enterprise needs. They have often chosen products based on existing vendor relationships without surveying the spectrum of available options. This situation is changing as market awareness and the scope of implementations grow.

Potential users are beginning to survey and weigh options more systematically, looking at criteria including:

- Industry- and application-specific experience and references; availability of application templates; customization and development possibilities;
- Language (linguistic) and feature identification and extraction support;
- Robustness and completeness of linguistic, statistical, and machine learning algorithms and the ability to extend the product;
- Supported file formats; ability to negotiate access to database and content-management systems;

- Data-exploration and output options – for interactive manipulation, visualization, and automation – that support organizational work practices;
- Ability to model and manage processes and to orchestrate them via external applications; integration with enterprise applications and analytical tools;
- Compliance with existing and emerging industry standards including frameworks such as IBM's Unstructured Information Management Architecture (UIMA) standard and established and emerging analytical methodologies;
- Vendor financial backgrounds and prospects;
- Vendor innovation and improvement track records;
- Licensing and support costs; and
- Need for consulting and professional services.

Users will want to benchmark and compare product performance and accuracy and other important characteristics under realistic volume and intake rate scenarios.

4.4 Deployment

Implementation best practices for text analytics, as for other computing disciplines, should focus on repeatable processes and verifiable outcomes, on performance, robustness, and correctness.

Best practices are facilitated by availability of industry- and application-specific process templates – essentially design patterns for text analysis – that guide the user in acquiring and processing documents, creating or adapting classification schemes, conducting analyses, and utilizing results. Adoption or adaptation of established methodologies can help. These may extend familiar data mining approaches such as:

- CRISP-DM, the Cross-Industry Standard Process for data mining;
- SEMMA, SAS Institute's Sample – Explore – Modify – Model – Assess knowledge-discovery process; and
- DMAIC, Define – Measure – Analyze – Improve – Control, process-improvement steps called for by Six Sigma methodologies.

Methodologies specifically focused on text analytics are emerging. ClearForest's ClearPath methodology, for instance, aims to unify text analytics and BI applications.

Flexibility is essential. Flexibility facilitates adoption of hybrid, semiautomated solutions that utilize human expertise for high-value steps and stresses reusability (including taxonomies and thesauri), and the ability to explain, act on, reproduce, and audit results.

4.5 Best Practices Conclusion

The creation of valuation-rooted best practices for the evaluation and deployment of text analytics, derived from theory and from user experiences and from familiar analytical methodologies, will facilitate organizational adoption of text analytics and integration of text analytics technology and processes into enterprise operations and analytical decision making.

Seth Grimes

Seth Grimes is a business intelligence, data warehousing, and decision systems expert, a consultant and contributing editor for *Intelligent Enterprise*. He writes the magazine's Breakthrough Analysis column and contributes feature articles.

Seth founded Washington DC-based Alta Plana Corporation in 1997 and consults on information systems strategy for clients that include primarily government statistical agencies and marketing firms and also select software publishers.

Seth writes and speaks on data management and analysis systems, industry trends, and emerging analytical technologies. He is chair of the Text Analytics Summit 2006 and also chaired the 2005 summit. His white paper for the 2005 summit, The Developing Text Mining Market, is available online at *http://altaplana.com/TheDevelopingTextMiningMarket.pdf*.

Seth can be reached at grimes@altaplana.com, +1 301-270-0795.

Text Analytics Summit 2006

Text Analytics Summit 2006 (http://www.textanalyticsnews.com), slated for June 22-23, 2006 in Boston, is a mindshare event for the leading developers, up-and-coming start-ups, tech-savvy users, and newcomers to the text-analytics space.

As the second year of the first commercially focused text-mining conference ever devised, Text Analytics Summit 2006 is an opportunity for vendors to identify the most promising applications, size up technical challenges, and connect with tech-savvy users eager to relate what they need. If you're a user of text analytics technology in any application or industry, this is an unmissable opportunity to learn from your peers and understand the bottom-line impact of the latest deployments. If you're a developer, this is your chance to meet a market focused on text analytics and exchange strategies for success with industry peers.

Alta Plana Corporation

7300 Willow Avenue
Takoma Park, MD 20912
+1 301-270-0795
altaplana.com