The Developing Text Mining Market

A white paper prepared for

Text Mining Summit 2005

Boston, June 7-8 2005

(www.TextMiningNews.com)

Seth Grimes

Alta Plana Corporation

Alta Plana

THE MARKET CHALLENGE

Text-mining technologies have won a strong but limited market beachhead by solving important business problems in a number of high-value application domains. They have proved their worth in diverse but narrow applications such as drug discovery, intelligence, and warranty-claims analysis. The diversity of these applications and the growing corporate understanding that a large proportion of enterprise knowledge is locked in text, that there are huge potential savings through automated text processing and profits through automated knowledge discovery, suggest that text-mining technologies are poised to break out into a much larger market. Packaging, positioning, and marketing strategy will be crucial in defining the extent of the market and in determining the speed, breadth, and depth of market up-take.

Text-mining vendors are fortunate that the existing and potential markets understand their context: everyone works with textual material in the form of contracts, e-mail, letters, memorandums, reports, and Web pages. We all understand that the information revolution of the last quarter century, impelled by PCs, e-mail, and the Web, has redefined employees, suppliers, sales and distribution channels, customers, and regulators into producers and consumers of information and not just of goods. Companies no longer simply buy materials, manufacture widgets, and sell products: they now consume and produce volumes of electronic information, the vast majority of it non-numeric, at each step along the way. Globalization and economic transformation toward services have accentuated this change.

Text-mining vendors are fortunate that, because their tools work on text rather than on numbers, they are not been to market based on algorithms. Where in the data-mining world you have to explain clustering, decisions trees, and neural networks, few potential text-mining customers will need or demand an explanation of support vector machines or singular value decomposition. Text mining can be sold as solving business problems without recourse to technical explanations.

Differentiation from search, which is low cost and easy to implement, is perhaps the biggest hurdle text-mining vendors face. Everyone knows Google and Yahoo!: they're easy-to-use and fast. Most users don't care that results are based on simple keyword search and often lack relevance. This complacency is evident in the reported simplicity of the composition of the vast majority of searches and the infrequency of users exploring beyond the first few returned hits.

The key text-mining differentiators are, of course, the ability to discover and organize knowledge, an ability that goes far beyond simply indexing documents and producing hit lists based on search terms. The key point is perhaps that the information content of a body of documents is far greater than that of a page examined in isolation. That information content may include not only words and phrases – search terms – but also entities such as names, e-mail addresses, phone numbers, chemical-compound names, and ticker symbols; concepts, which essentially place terms and entities in a domain-based context; and especially interrelationships, whether concept- or entity-based or temporal or other.

Business, academic, and government markets are primed for the introduction of more sophisticated knowledge tools such as text-mining software that can advance an organization beyond content management to knowledge analytics. As we will see, tool, solution, and service vendors will do well to craft marketing strategies with regard to application function, targeted business domain(s), and integration with operational/transaction systems.

1 Perceptions and Positioning

Understanding market perceptions within the text-mining arena and in areas related to text mining, whether because their technologies compete with or complement text-mining technologies or because they only appear to, is essential in understanding and shaping the developing text-mining market and to optimal market positioning.

Does text mining fit best alongside data mining in a technically focused knowledge discovery category? Or is text mining best seen as an extension of content and knowledge management systems that classify, store, and deliver documents? Or does text mining contribute most when integrated into operational systems such as customer relationship management systems? Of course the answer is Yes to each of these questions at different times and to different audiences. It is our job to understand those times and audiences.

We will start by looking at characterizations of text-mining and related technologies.

1.1 Knowledge Discovery

Text mining is a specialization and extension of the broader data-mining field, a form of knowledge discovery. The source information for text mining is documents versus fielded, numeric data so we apply linguistic techniques to create machine-processable meaning and tag document elements such as entities and concepts. These steps take the place of the data preparation one does with data mining to make the data susceptible to analysis. In the two knowledge-discovery variants, once the data is prepared – once document elements are tagged or numeric data is put in well-formulated data structures – there is a similar application of statistical techniques to reduce problem dimensionality, to identify correlations, clusters, linkages, and relationships, and to create predictive rules, also known in some circles as knowledge.

The commonality of the statistical knowledge-discovery techniques is well worth emphasizing as it correctly and necessarily establishes text mining's well-foundedness. That the starting point for text mining is sets of readily comprehended textual documents is a boon to the acceptance of text mining by IT and business specialists who may have a suspicion of numbers that resist easy understanding. Text mining's outputs are another comparative asset to acceptance as they are far more easily comprehended than the often obscure models and parameters created in mining numeric data.

1.2 "Unstructured Data"

Text is generally grouped in an "unstructured data" category along with videos, images, audio, and other information stored with little or no internal subdivision into fields with constrained, machine-recognizable meanings or use of data types other than undifferentiated binary objects. Certain types of textual documents, those based on forms or marked-up with XML, are termed "semi-structured" where the forms' field labels and the XML tags provide some degree of machine-processable semantics.

The "unstructured" label groups text and other forms in contrast with fielded, numeric, database-managed data. For text, however, that label is inaccurate and inapt and paints an incomplete picture of what text mining can accomplish.

First, textual documents have linguistic structure that is easily comprehended *by people* on multiple levels. Individual words have declensions (nouns) and conjugations (verbs) and, in some languages, cases that indicate the thematic or semantic role of a phrase in a sentence.

Sentences have grammatical structure, built from phrases with subject, verb, and object and auxiliaries such as prepositions. Paragraphs, chapters, etc. build on smaller structural elements and provide contextually-derivable meaning to elements at each level. Text-mining tools apply linguistic (natural language) techniques to parse and model the structure inherent in textual documents.

Secondly, the "unstructured" label is inapt in that it minimizes the differences in handling visual images and textual documents. The modeling and pattern-recognition approaches are significantly different in application even if not in theory. And customers would apply text-mining and, for instance, image-mining solutions for very different situations.

Lastly, the "unstructured" label implies that creating structure is the goal. It is not. The goal is to discover knowledge – categorizations, relationships, predictive rules – and apply the discovered knowledge in document processing via clustering, routing, and analysis on extracted information. The "unstructured" label diverts attention from the knowledge-discovery goal to the characteristics of the source information.

While the "unstructured" label is unlikely to go away, it's use might best be minimized.

1.3 Content, Search, and Analysis

Text-mining technology interacts and overlaps with tools and solutions in a number of market-established categories. To understand the evolution of text-mining market, we must understand how the technology has been and can be positioned *vis à vis* –

- Content Management, Knowledge Management, and Content Delivery (Portals)

  Text mining can and should be rightly be seen as adding value to content management (CM) and portal systems. CM provides, essentially, a combined historical-operational document repository but focuses on management and retrieval rather than on knowledge discovery. In industry/vendor usage, knowledge management (KM) is essentially synonymous with CM albeit with an added facet, business rule- and metadata-management. Portals often serve as CM/KM user interfaces, providing a mechanism for search and document delivery without providing important analytical capabilities that would allow user-initiated extraction of relationships and rules from document stores.

- Enterprise, Desktop, and Web Search

  Enterprise search is the ability to index documents in disparate formats, residing in repositories, databases, and file systems across an organization, for retrieval based on search terms and on properties such as creation and last-edit data, document format, author, and subject classification. Enterprise search tools are expected to negotiate with rights-management systems and usage restrictions. They are often tightly integrated with CM systems and incorporate significant text-mining features.

  Desktop search is typically much less than a personal version of enterprise search: most implementations search only file systems and provide little or no relevance ranking although they often do provide nice visualizations. The desktop-search picture is changing however as Web-search tools such as Google are scaled down to the desktop. The promise is to replicate the user's Web experience in searches of her own local files and e-mail.

  Text mining shares with enterprise search the challenge of differentiating from Web search. Google, Yahoo, and lesser competitors are ubiquitous, simple, powerful, and

cheap. They are recognized consumer brands that have crossed over into the business world. And they seemingly cover the same territory as as enterprise search and text-mining tools, making Web search perhaps the most significant competitor these latter tools face.

- Business Intelligence

  Business intelligence (BI) encompasses a set of technologies and techniques whose compass differs for different definers. Minimally BI includes query, reporting, and dimensional analysis (namely, OLAP, On-Line Analytical Processing) of numerical data that resides in data warehouses and topic- or usage-specific data marts. Some would expand BI to include infrastructure technologies for ETL (extract, transform, load), data cleansing, data integration, metadata management, and data warehousing. BI's growth areas are business performance management and operationally embedded analytics. Nonetheless, analytical reporting, with the addition of dashboard displays of derived indicator variables, has been and will remain BI's bread-and-butter application, and Microsoft Excel, with all its limitations, the most used analytical software tool.

  BI vendors focus on numerical data and have largely ignored text-mining as alien to their world. There are exceptions, BI vendors whose most important products are database management systems and statistical systems, enterprise applications, and industry verticals.

- Visualization

  Visualization is an excellent complement to knowledge discovery, an integrated component of several leading text- (and data-) mining systems. Innovative, interactive displays are especially useful for navigating hierarchically structured document sets and for exploring concept- and entity-derived relationships.

- Enterprise Applications

  Enterprise applications from vendors such as SAP, Oracle, Peoplesoft, and Siebel automate business operations by providing interfaces and data structures for human resource management, financials, manufacturing, logistics, marketing and sales automation, customer relationship management (CRM), and other key corporate functions, all built on customizable, best-practices business-process management. Each of the functions covered generates and consumes huge volumes of data. Building text-mining into enterprise applications would augment value to users.

1.4 Technical Approaches and Differentiators

In application, text mining starts with lexical analysis of a document: stemming words, performing frequency counts, and determining roles and associations. Text-mining tools extract, tag, and analyze associations among identified entities and concepts and the documents that contain them. They create categories or they may apply existing taxonomies – hierarchical knowledge representations – to classify documents, and extracted data may be used for other forms of analysis. They apply statistical techniques to cluster documents according to user-determined characteristics. Lastly they deliver both interactive exploratory capabilities and hooks to allow classification to be embedded in applications to add automated text processing.

The ability to stem words, identify phrases, and extract terms and entities is shared in

degrees by search tools, which are however built for document retrieval rather than analysis and exploration of document sets, statistical-analysis functions that are not part of Web search or content management offerings. Knowledge discovery – pattern recognition – via application of both statistical and linguistic techniques is a key differentiator of text mining from those latter technologies.

Because text mining looks at document sets and identifies inter-document relationships, it supplies context that enables far greater relevance to search results than is provided by search tools. Contextual relevance – the ability to apply domain knowledge to match patterns and cluster results – is a second key technology differentiator.

Lastly, text mining tools can be embedded in applications that produce and consume significant amounts of textual data and often pose real-time operational demands. Content-management and enterprise-search tools do not offer the same potential for operational integration.

1.5 Category Confusion

It is standard practice for most information-technology vendors to expend significant effort defining their technologies and solutions and positioning for industry analysts. It is deemed important to occupy the rarefied "magic quadrant" reserved for visionaries with high ability to execute. While a position toward the upper right of an analyst's scattergram is unquestionably a marketing asset, text mining appears to be an IT specialization where analyst firms have miscategorized the technology and where for certain vendors, assessed ability to execute is derived from product lines other than their text offerings.

Leading analyst firms seem to prefer to group text mining with content and knowledge management, in essence overemphasizing aspects that relate to knowledge extraction (e.g., generation of taxonomies and classifications) and ties to document-management systems and portals and denigrating the value added by analytics. The reason for analyst grouping of text mining with content management is likely that the dollar volume and extent of the CM market is far larger than those of the data-mining market.

Classifying most text-mining tools in the content management and delivery category is akin to classifying a business intelligence suite like Microstrategy's as a database system.

The focal point for business intelligence is shifting from reflective analysis of historical data to the operational application of model-based analytics for "right-time" business decision making. Similarly, the greatest long-term value of text mining is likely to be in operationalizing features such as entity and concept extraction and automated classification and knowledge-based predictive analytics: areas of kinship with data mining. It would be constructive to educate analysts working in data-mining about text mining's ability to extend predictive analytics by crunching textual information.

2 MARKET SEGMENTATION

It is instructive to divvy up the text-mining market by the type of customer. Applying a bit of Bayesian reasoning, early buyers (as text mining is a relatively new technology) will prove a good indication of market directions. Nonetheless, text mining is testing new directions as the technology has made possible several new applications.

2.1 TECHNOLOGY BUYERS

Technology buyers are heavily represented among early adopters. These are individuals and organizations who have exacting needs for accuracy, speed, volume, and automation accompanied by the willingness to bet on emerging technologies. Intelligence and law-enforcement agencies with a counter-terrorism mandate fit squarely in this market segment. They need to sift massive volumes of text, handling multiple languages and dialects, extracting names and other identifiers with variant spellings and also events, to make associations and detect geospatial, temporal, and inter-personal relationships. They need to do this quickly and to produce convincing, actionable results.

Technology buyers demand or roll their own applications tightly focused on particular business domains and functions. These solutions may often be generalized to create solutions for other business domains and for for other business applications.

2.2 APPLICATIONS AND VERTICALS

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 mining is long. It includes –

- Chemical/Drug/Gene/Symptom Discovery

  Sources include scientific and clinical literature, treatment records and reports, databases of chemical compounds, phamaceuticals, and medical symptoms, and so on.

  The goal include to extract relationships among biomedical and chemical entities and 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."

  Entity extraction here is recognition of 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 Counter-Terrorism

  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 inter-personal and organizational) and support of criminal investigations and prosecutions.

- Securities Fraud Detection

  Sources include financial & news reports, corporate filings & documents, trading and other transaction records

  The goals include detecting insider trading, reporting irregularities, money laundering and illegal transactions, and pricing anomalies, etc.

... and legal discovery and strategy development; patent examination; recruitment 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 mining can play a significant role in may 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 mining 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-mining 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 tools.

2.3 PLATFORMS

Some organizations are platform buyers: they implement packaged enterprise-scale solutions and are willing to try capabilities that come integrated with those solutions. Notable platform examples include ERP systems such as SAP, Siebel, and Peoplesoft with integrated vendor-provided and third-party analytics; database systems such as Oracle, IBM DB2 and Informix, and Microsoft SQL Server with built-in OLAP engines and object extensions that include, for Oracle and IBM, text management; and business intelligence suites from vendors such as Business Objects, Cognos, Hyperion, Microstrategy, SAS, and SPSS. These examples all constitute platforms that can and are or will be extended with data- and text-mining capabilities, in many cases licensed from specialized third-party vendors.

3 Go-To-Market Strategy

The text-mining market is relatively new and not yet rigidly defined. Growth potential is huge given the ever-increasing volumes of textual information being produced and consumed by industry and government. While the market is being targeted by established software vendors, which are extending existing product lines and existing functional and industry-specific offerings, several pure-play vendors have done quite well and entry barriers for innovative start-ups are still not high.

Given the number and varied nature of market opportunities, vendors must assess market developments frequently and refine their packaging, positioning, and marketing strategies accordingly. The first question is, What are you selling? The next is perhaps, What companies are you competing against whether head-to-head or indirectly? A third area to explore is alliances that would exploit technology and business syneries.

3.1 Product Definition: Technology or Solution?

Current text-mining offerings are for the most part toolkits with interfaces and, often, industry-specific modules although there are notable exceptions of operational applications that embed text-mining capabilities to support automated classification. Text mining has been positioned primarily as a technology rather than as a solution, and where it has been packaged as a solution, that packaging has been as part of larger solution.

3.2 the Competitive Landscape and Technology Channels

It should go without saying that the greatest competition for any given text-mining vendor comes from other text-mining vendors, which include pure-play companies, data-mining vendors, and certain platform vendors. The existing market is far from saturated however, and the potential market is large. But this white paper is vendor-neutral and not intended to provide advice for any particular vendor in competing with others so the author will beg off any discussion of vendors in or aspiring to the text-mining market. Other competitors and potential collaborators are found in a number of areas that include –

- Web Search

  Free and easy Web search, with tools extending their reach into enterprises and down to the desktop, is the biggest market inhibitor for text mining. Search – finding and retrieving documents – is used by everyone who uses the Web, and it is rapidly evolving, for instance with the addition of results clustering with visual navigation of result sets. Yet while the major search presences use knowledge discovery techniques, their business models are aimed at increasing distribution of context-relevant advertising rather than at text-enabling the spectrum of enterprise computing applications. The Web search vendors will not gobble up text mining's potential market. Rather they may themselves present a market for text-mining vendors.

- Enterprise Search

  Enterprise search vendors constitute text mining's closest competition from a technology point of view. But unlike most text-mining vendors, they are selling information-delivery solutions rather than analytical technologies, leaving wide open the market for tool-oriented text-mining vendors. Similarly these vendors haven't touched analytically and numerically intensive applications. They do not compete with text-mining vendors whose products are integrated with data-mining and other data-analysis products.

- Data Mining and Business Intelligence

  There are a number of data-mining vendors whose offerings center on predictive models for credit rating, risk analysis, fraud detection, and the like, areas where inclusion of data from textual sources would be a huge bonus. While several prominent data-mining vendors have their own text-mining tools, one of the most notable licenses software for linguistic analysis from another vendor. Data mining and business intelligence vendors provide complementary analytical products and are potential business partners more than they are text-mining competitors.

- Enterprise Applications

  Enterprise application vendors do not currently provide text-mining capabilities to any *significant* extent although the CRM vendors in particular are moving to add these capabilities to their products.

3.3 Business Strategies and Synergies

In addition to the technology alliances already described, there are other business synergies that if exploited would win market space for text-mining vendors.

- Development platforms

Software development platforms provide interface-development environments (IDEs) and provide frameworks for plugging in code modules. The most notable development platforms are the open-source Eclipse IDE (Java) and Microsoft's Visual Studio (.Net).

- Services Model

Many smaller technology providers have found that releasing products to open-source developers (who may include their own employees) is an effective way to encourage rapid market up-take and accelerate development while providing an excellent revenue stream from services and support.

- Information Producers and Distributors

Many, many organizations produce, aggregate, and distribute information: news services; financial-data wires; scientific and business journals; print and electronic publishers; survey researchers; and other data producers. The value of their offerings could be greatly enhanced by the additional of user-driven text analytics and visualization of text-extracted data and knowledge.

- Application Service Providers

Text mining could be a useful addition to the services offered by application service providers (ASPs), organizations that provide hosted applications or that mediate transactions, the latter including trading exchanges.

4 CONCLUSIONS

While text mining has achieved clear successes in noteworthy high-value applications, there are obstacles to success in the broad market and across the enterprise. The most significant obstacle is that the technology and its value are not yet well understood in comparison to technologies and services that are similar but better established and far better know, namely content management and enterprise and Web search. Effective packaging, positioning, and marketing can overcome this obstacle.

Key elements of successful product-development and marketing efforts will include –

- Differentiation from search via strong communication of the value of knowledge discovery and context-sensitive, relevant information retrieval that goes far beyond the ability of search engines to find and list document hits.
- Differentiation from content management via strong emphasis on text analytics, on text mining's integration with data mining technology that establishes relationship, classifies, and clusters documents according to extracted concepts and entities and detected associations.
- Communication of the ability to operationalize text mining via integration with customer relationship management, marketing and sales automation, manufacturing, human resources management, and other enterprise applications, an aim that would be furthered by technology partnership with leading and niche vendors.
- Exploitation of early successes in diverse verticals such as drug discovery, health-care review, warranty-claims analysis, etc. to create new verticals.
- Creation of new solution and service categories such as Relationship Management and Reputation Management that could not exist without text mining's automated knowledge-discovery capabilities.
- Integration with business intelligence products applied at levels from tactical/ departmental to strategic/enterprise and used for both exploratory and operational analysis.
- Emphasis on usability, on the output side by delivering results in appealing dynamic, graphical forms that support interactive exploration of document sets and discovered concepts, associations, and relationships, and on the input side by integrating text mining with established data and content management solutions to create an enhanced text-processing platform.

Business, academic, and government markets are primed for wider introduction of text-mining as a knowledge tool for reflective and predictive analysis and as an integrated component of key enterprise applications. Packaging, positioning, and marketing strategy will be crucial in defining the extent of the market and in determining the speed, breadth, and depth of market up-take.

Alta Plana Corporation

Alta Plana's business is information systems architecture and development. Starting from the premise that technological choices should derive from business requirements and organizational dynamics, we help organizations design and implement workable strategies that incorporate *appropriate* modern technologies.

Alta Plana specializes in data management, analysis, and dissemination technologies: data interchange and integration software, database management systems, and analytical and decision support solutions. Alta Plana staff have extensive experience working with demographic, economic, social, marketing, and engineering data and statistics and with associated metadata and business processes.

Alta Plana works for governmental organizations, technology and solutions vendors, and commercial and non-profit corporations.

Seth Grimes

Seth Grimes is a business intelligence, data warehousing, and decision systems expert, a consultant and an Intelligent Enterprise magazine contributing editor and writer, author of the magazine's Breakthrough Analysis column. Seth founded Washington DC-based Alta Plana Corporation in 1997 and consults, writes, and speaks on data management and analysis systems, industry trends, and emerging analytical technologies.

Seth can be reached at grimes@altaplana.com, 301-873-8225.

Text Mining Summit 2005

Text Mining Summit 2005 (http://www.textminingnews.com), slated for June 7-8 in Boston, is a mindshare event for the leading developers, up-and-coming start-ups, tech-savvy users, and newcomers to the text-mining space. As the first commercially focused text-mining conference ever devised, Text Mining Summit 2005 is an opportunity to identify the most promising applications, size up the technical challenges, and connect with the tech-savvy users eager to relate what they need.

Alta Plana Corporation

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