What's Next for Text

Text Analytics Today and Tomorrow: Market, Technology, and Trends

A white paper prepared for

Text Analytics Summit 2007

Boston, June 12-13, 2007

www.TextAnalyticsNews.com/usa

Seth Grimes

Alta Plana Corporation

Alta Plana

Table of Contents

1 INTRODUCTION

Text analytics has entered the IT mainstream:

- The technology has matured in ways that reduce complexity, boost accessibility and usability, and widen appeal for end users.
- Usage has broadened to encompass new types of applications and new business domains.
- The field is supported by a vibrant, international research community with strong academic, government, and industry participation.
- Sales are strong and growing at a rapid pace with available options ranging from a wealth of specialized, free, open-source software tools to enterprise scale, integrated analytical systems.
- The market outlook is creating opportunity for new entrants and prompting a healthy level of merger and acquisition activity.
- The value of text analytics and its potential enterprise role are newly understood by the wider world of analysts, trade journalists, and end users.

Text analytics' successes date back up to ten years, achieved in fields such as life sciences and counterterrorism. The technology has notably revolutionized the pharmaceutical drug-discovery process. Scientists can mine biomedical literature for protein interactions that suggest therapies, reserving time-consuming and expensive laboratory work and clinical trials for the few, most promising leads. Intelligence analysts can similarly use automated methods to discover telling patterns hidden in large volumes of textual materials, materials that were previously inaccessible due to source language, volume, and resource constraints.

Yet investigative analyses have been pushed from the text-analytics center stage. While many users still seek needles in haystacks, more-recent text-analytics adopters are most interested in statistically described characteristics of larger populations. They're looking for trends and big picture rules rather than for the few exceptional cases. They want a matching style of analysis, one that has served them well in working with conventional, numerical data sources. They want familiar business intelligence (BI) interfaces and techniques, also adaptation of concepts such as Extract, Transform, Load (ETL) to their work with textual sources. And they want software that deals automatically with linguistic complexity and delivers information in the structures and interfaces they're using for line-of-business applications and analytics.

The shift from a knowledge-discovery to an analytics paradigm is part of text mining's expansion to support a larger world of enterprise analytics. This shift allows and encourages a majority of organizations to investigate and adopt the technology. The questions of those concerned with enterprise analytics are no longer *if*, *what*, or *how*, but rather *where* and *when* – questions this paper will address in examining *What's Next for Text*.

2 Market Success

> "Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before.[1]"
>
> *– Philip Russom, the Data Warehousing Institute*

2.1 From Mining to Analytics

Text mining derives from data mining, a statistically rooted approach to classification, clustering, and derivation of association rules, with data mining applied to fielded, numerical data and text mining extending processes to encompass textual data sources. The aim is description and prediction. The technologies are typically embedded in larger knowledge-discovery efforts that add data management and visualization to the mix. But where semantics in data mining – the meanings of variables and values – are captured in variable names and descriptions and other metadata, when working with text, it's a series of analytical steps that infers machine-comprehensible semantic meaning in "unstructured" textual source materials.

Text is replete with linguistic and grammatical and compositional structure – the form and arrangement of the words and sentences from which documents are built. People exploit this inherent structure without conscious thought when they read. Machines must similarly exploit it to extend automated knowledge discovery to text. They use analytical software that annotates and restructures text into mineable form via small but important steps that discern the morphology of words, the boundaries of sentences and other syntactical units, and the contextual meaning of terms.

Search is one of the Web's (and the enterprise's) killer apps, and ironically search's short-comings suggest the value text analytics can deliver. Users want to work with natural language and they want more than the information retrieval – essentially, document retrieval – that search (currently) offers. Text analytics add semantics possibilities both in the processing of search queries and in the use of search findings. Secondly, search is poised to transform the business intelligence world, again aided by a dose of text analytics. Vendors have struggled for years to make BI accessible to a greater variety of end users and for a greater variety of applications. Operational BI, dashboards and portals, visualization, and spreadsheet integration have all advanced BI technology toward meeting this goal, but it is the promise of cutting through the clutter of menus and catalogs that is the best candidate to deliver.

2.2 From Tools to Solutions

A shift in market focus from tools to solutions is a hallmark of market maturation. Tools work well for technically oriented researchers – traditional text miners – who expect to do significant configuration and customization in order to accomplish highly specialized tasks. This class of users works with software components and workbench interfaces to design processes that meet individual needs. Such tool-centric approaches serve them well but are unsuitable for the broader world of enterprise data and business analysts whose needs are solution oriented.

One can build solutions quickly when software components are easy to configure to

[1] *www.tdwi.org/Research/display.aspx?id=8355*

meet domain- or application-specific needs while providing a high level of capability and performance for generalized problems. The market has evolved in recent years to provide such solutions. Text analytics is now enterprise-ready:

- Entity extraction accuracy – precision and recall – is high enough to meet the needs of a broad variety of applications in disparate business domains.
- *Usable* solutions are now available to address the challenge of tapping the axiomatic 80 percent of enterprise information that is locked in textual form.
- Packages are available for mainstream business needs – survey processing, competitive intelligence, CRM, reputation management – that integrate with line-of-business applications and conventional analytical software.

Therefore:

- Text analytics is now accessible to business analysts, not just linguists and statisticians and scientific researchers.

And enterprise integration is facilitated by widening Unstructured Information Management Architecture (UIMA) compliance. The specification for this open-source, platform-independent, software-integration framework is being developed for "interoperability, discovery, and composition of analytics across modalities, domain models, and frameworks"[2] by an OASIS technical committee, and creation of a reference implementation is now an Apache incubator project.

2.3 Emerging Applications

> Growth is largely driven by the wealth of unstructured information found on the external web, in corporate intranets, document repositories, call-centers, and in customer and employee business communications.[3]
>
> *– UIMA Chief Architect David Ferrucci and fellow IBM researchers*

The earliest text analytics successes have been in investigative applications in the life sciences and intelligence. Breck Baldwin of Alias-I goes so far as to state that "the Nobel Prize in Medicine will [soon] go to a researcher primarily focused on text analytics because of the discoveries made possible by the technology." Such segments will remain important, but the majority of market growth will occur elsewhere.

Justin Langseth, president and cofounder of text-mining vendor Clarabridge, offers that "the biggest changes over the last year have been squarely associated with extremely rapid growth of the text-mining market within the commercial sector, primarily for marketing uses – product positioning before launches, competitive intelligence, quantifying sentiment, early issue detection – and for customer service and product management uses – listening to the voice of the customer to better understand their needs, frustrations, suggestions, etc."

Emerging applications include market research and survey analysis, reputation management via monitoring of networked social media, and semantic enrichment of Web and enterprise search. We will also look at extension of analytical approaches to media beyond text. These will be important areas for text analytics in the near term.

---

[2] *xml.coverpages.org/UIMA-TC-Announce.html*

3 "Towards an Interoperability Standard for Text and Multi-Modal Analytics," *domino.research.ibm.com/library/cyberdig.nsf/papers/1898F3F640FEF47E8525723C00551250/$File/rc24122.pdf*

2.3.1 Sentiment Extraction, Market Research, and Survey Analysis

For market researchers and survey analysts, text analytics is a compelling replacement for slow, expensive, and cumbersome old-school approaches to dealing with free-form information. One human-resources manager interviewed reported that text-analytics software allows him to accomplish in half a day work that formerly took eight analysts one week. Further, automated processes may be integrated with line-of-business software, for instance for Customer Relationship Management. This integration creates a new market category, Enterprise Feedback Management.

These applications rely on sentiment-discovery capabilities that go beyond basic entity extraction. Sentiment can be hard to discern and quantify accurately, so Anne Milley, a SAS Institute marketing executive, advocates a hybrid approach that joins fielded ratings with text analysis. Milley states that "model accuracy is good – and we believe most likely better than models that use non-supervised methods."

Reputation management further applies sentiment extraction and related techniques to social media: blogs, message boards, and other online sites where opinion can propagate quickly. Organizations wish to detect postings that cite them and their products, competitors, and customers. They wish to handle incidents expeditiously to maintain a positive image, and they rely ever increasingly on text analytics.

2.3.2 Search Enrichment

Users tolerate often-low search accuracy because consumer search is free and far better than inability to find information. One tactic, manual tagging based on *folksonomies*, improves findability and retrieval relevance but can be only a partial remedy due to the Internet's volume and volatility.

Where established and informal search and tagging methods fall short is where semantic technologies can play a leading role. Semantic indexing and search provide the possibility of going beyond keywords to exploit intent, context, and concepts inferred from both search queries and target documents.

Enhanced search is an excellent resource for investigative analyses and may obviate the need, in many cases, to extract information to databases. Alias-I's Baldwin explains that "database filling will no longer be seen as the end goal of high-end text analytics. Reversing a long standing trend, enhanced search portals are the new focus of technologies ranging from brand sentiment to entity and relation extraction."

2.3.3 Media Mining

User organizations and vendors recognize that while there's a lot of text out there to tap, there's more to "unstructured" information than just text. In terms of technology application, free text is simply a nearest next step away from fielded data into a world of diverse media. Media miners automate information extraction from speech, sound, still images, and video and the analysis and processing of those forms. Important early applications are the same as those text analysts tackle: automated handling of customer communications, security applications, and the like, but with a need to apply native methods, for instance the use of phonology to mine speech rather than just blanket transcription to text. There are dissimilarities, for instance, media mining may involve a real-time angle that is not imperative in text processing. Nonetheless, integrated data-text-media analytics is another area that is *next for text*.

3 Continuing Challenges

> Went finding and seeking,
> Finding less than sought
> Seeking more than found,
> Every detail minding
> Of the seeking or the finding.
>
> – *Langston Hughes,* Old Walt

Text analytics meets many challenges; nonetheless, we seek ever more powerful technology to respond to evolving demand. Many challenges are unmet or partly met. They suggest w*hat's next for text* and are well worth exploring.

3.1 Question Answering: Search Done Right

Few search users are looking for the hit lists of dubious relevance returned by the major search engines. Rather, most searches are really questions – *Who is the president of Peru? What's the best price in the Washington, DC area for a Novara Safari bike?* – and most searchers are seeking facts rather than undigested documents. It is text analytics that will deliver the capabilities needed to provide answers, for search done right.

Any attempt at question-answering involves understanding the searcher's intended meaning, finding suitable information sources, extracting facts, and composing responses. Matthew Glotzbach, head of products for Google Enterprise, has said that "question answering is the future of enterprise search," a statement that applies equally to public Web search. Advanced text technologies are an essential ingredient.

3.2 Multilingualism

Multilingual and cross-lingual text analytics – going beyond *multiply monolingual* processing – is perhaps more talked about than usable. We wish to exploit text without regard to language, which requires more than ability to operate in a single language or to translate from foreign languages into a canonical language, whether English or some other. Software must understand, equivalence, and jointly process vocabulary, grammar, and concepts for disparate languages. These functions are nominally within the capability of many existing text-analytics tools, yet for most users, real-world constraints come into play. In the words of an executive of one multilingual text-analytics vendor, "Everyone seems to ask about it, and they often have a checkbox on their RFPs for it, but in practice, dealing with the massive amount of data in English first keeps foreign language handling off the table."

3.3 Semantic Web Services

The term *semantic Web services* could refer both to services for the still-elusive Semantic Web and to Web services that provide semantic analysis on-demand to invoking applications. The first referent is a vision whose realization remains elusive a decade after it was first articulated by Web inventor Tim Berners-Lee: it is difficult to get information publishers to mark up their content. The second would provide an automated complement to *folksonomies* that overlay semantics on the existing, chaotic computing world: think semantic *mash-ups*. Initial attempts at this second type of semantic Web services, albeit short on accuracy, are available. They are an important step in another direction that text analytics is heading.

3.4 A Grand Challenge

Ronen Feldman, professor of computer science at Bar-Ilan University and cofounder of text-analytics vendor ClearForest, proposes a "grand challenge" for text mining[4]. Feldman's grand challenge is expressed in testable terms:

> Text mining systems that will be able to pass standard reading comprehension tests such as SAT, GRE, GMAT, etc.

Meeting this grand challenge entails improved Entity Extraction,

> We are seeking domain independent and language independent [named entity recognition and] relation extraction systems that will be able to reach precision of 98-100% and recall of 95-100%. Since the systems should work in any domain, they must be totally autonomous and require no human intervention.

Those Autonomous Text Analysis systems

> will analyze huge corpuses and come up with truly interesting findings that are not captured by any single document in the corpus and are not known before... Such systems can then be used for alerting purposes in the financial domain, the anti-terror domain, the biomedical domain and many other commercial domains.

Feldman believes that within five years researchers will be able to demonstrate systems that meet the Extraction and Autonomicity targets, using the Web (and presumably other corpora) for above-average scores on grand-challenge test questions.

Yet the best-of-five answer to a multiple-choice question, based on a few paragraphs of clean text, could conceivably be picked by moderately sophisticated pattern-matching software. Developers are more ambitious, working to boost accuracy when dealing with real-world information sources and conditions, to deliver the ability to mine noisy materials such as call-center notes, survey responses, e-mail, and the like. Here, syntax may be fractured and ungrammatical, spelling is irregular with abbreviations, and a given source document may contain externalities, what a linguist would call exophora, or references that are not resolved in examining a single source document. And, noting that much information found on the Web (and in the enterprise) is of dubious accuracy and authority, a next-generation, high accuracy, real-world, autonomous text miner would assess and weigh the correctness of identified responses in order to formulate a single, contextually best answer.

Feldman cites the Turing Test as, essentially, a generalization of his text-mining grand challenge. The Turing Test is a conversation: Can a person tell that an interlocutor is a machine? Conversations take place over time. They flow and meander and are sometimes discontinuous. The language usage is fractured and meaning may depend on context and externalities. And responses are contextual rather than selected from among a small set of prepared choices in an artificial test. Passing the 57-year-old Turing Test, suitably updated, may be the best real-world grand challenge, guiding researchers toward a comprehensive treatment of text.

Text analytics is already part-way there: witness visualizations of mined information ranging from networks of protein-protein interactions to the rendering of search-engine results in hierarchical clusters. Yet many essential tasks remain undone. Accomplishing those tasks in the coming years is *What's Next for Text*.

---

[4] "What Are The Grand Challenges for Data Mining? KDD-2006 Panel Report," *www.acm.org/sigs/sigkdd/explorations/issues/8-2-2006-12/9-is-there-grand-challenge-for-dm-explorations-final.pdf*

4 Market Outlook

The market boasts a diversity of text-analytics products that accommodate a spectrum of application types and business domains. The market includes pure-play text analytics and computational-linguistics vendors. It includes solution providers that target particular applications and domains and it includes companies offering integrated analytics. Search vendors loom large, and business-intelligence companies are looking to add text-analytics to their product sets. Content-management vendors are involved although their interest is largely limited to information retrieval.

> The author estimates a worldwide market for text-analytics software licenses, support, and professional services of about $200 million for 2006 with 25 percent annual growth through 2010.

This figure covers pure-plays, targeted solutions, and an allotment of the portion of other-analytics revenue that is attributable to text analytics. By contrast, Gartner estimates worldwide business-intelligence software licensing, exclusive of services, at around $2.5 billion with growth under 10 percent. Applying a conservative multiplier to compute user/contractor labor, and assessing the value of academic research and of content and database products produced with technology developed in-house by publishers, we arrive at a $2 billion valuation of worldwide text-analytics research, software, services, and applications.

4.1 Vendor Trends

The commercial landscape reveals both stability and opportunity, by-products of the rapid pace of growth in adoption and interest. Market development will continue accordingly with further consolidation, clarification, alliances, and new entrants.

4.1.1 Consolidation

Text analytics is part of a larger data- and content-management analytics market. Vendor consolidation must be considered in context of the larger market. The deals are significant even if smaller than, say, Oracle's $3 billion purchase of business-intelligence vendor Hyperion. The largest text-technologies transaction was the late-2005 Autonomy acquisition of enterprise-search rival Verity, Inc. for approximately $500 million. The combined company has annual revenue of over $200 million. More recently, BI/integration vendor Informatica paid $55 million for Itemfield to add the ability to tap "unstructured" sources to its ETL suite. The deal closed in December, 2006. And enterprise-search giant Fast Search & Transfer has announced a $23 million purchase of Convera's RetrievalWare business unit that is expected to close in the second quarter of 2007. RetrievalWare provides context-aware information access using a variety of text-analytics technologies that will complement current FAST capabilities.

Expect a quickened pace of merger and acquisition activity as database, BI, and enterprise-applications vendors seek to add text technologies to their product lines. Smaller companies, including some that are struggling, are particularly inviting targets.

4.1.2 Clarification

There is significant room for market clarification due to the variety of technologies in

use and the ways vendors have positioned themselves. On the technology front, we have traditional inverted keyword indexes for search, varieties of semantic indexing and analysis, Bayesian supervised classifiers, and relational-linguistic entity extraction. As vendors consolidate – as in the Autonomy-Verity merger – and as they add capabilities, the market will require technology clarification.

That clarification may reinforce product categories or it may blur boundaries and reduce them. Products already sit in multiple categories, such as computational linguistics, text-analytics companies, general/integrated or industry analytics, text-BI, Web and social-media mining, contextual advertising, content database publishers, and search vendors. These categories are useful for product differentiation but may hinder maximum market growth

4.1.3 Alliance

The text-analytics market has many inter-vendor alliances. Many solution providers, both within the text-analytics world and in nearby content-management and BI segments, license entity-extraction and linguistic technologies for resale in "OEM" partnerships. Expect such cross-licensing arrangements to continue due to the specialized nature of certain general-application technologies.

4.1.4 Emergence

Text analytics is fertile ground for new offerings. Most often, new products and solutions emerge as the commercialization of academic research or as the product-realization of in-house industrial research. Much basic research is government funded in the U.S. and by the European Commission, particularly when it addresses scientific, linguistic, and homeland-security challenges. Other, more immediately commercializable research is venture funded.

In other cases, such as SAIC's January 2005 sale of its Content Analyst division and intellectual property, new products emerge as spin-offs.

4.2 Market Validation

Consider text analytics as an industry and not just a technology. Analyst attention and willingness of funders to make new and repeat investments are strong validation of text analytics' vitality and potential.

Analyst firms have newly added text analytics to their research portfolios, a success indicator that validates the importance and commercial value of the market. Whereas in past years only IDC and the 451 Group tracked and reported on text analytics, analysts from Forrester, Gartner, Hurwitz, and the Data Warehousing Institute have joined the fray. Look for more in-depth and frequent analyst coverage as the firms respond to a growing volume of client requests.

And in the last year, venture funders have placed multiple, multi-million dollar investments in companies including Clarabridge ($7.2 million), InforSense ($10 million), and Nstein (CAN$10 million). Growth of companies including Attensity, Basis Technology, ClearForest, Inxight, SRA (NetOwl), and TEMIS has been fueled by strategic infusions of venture capital. Investors range from conventional venture firms to organizations such as IBM and In-Q-Tel that actively promote leading edge technologies to economic-development agencies such as ITI Life Sciences. Look for continued capital investments to support new entrants and growing firms.

5 Appendix: Technology Basics

> Thus the orb he roamed
> With narrow search, and with inspection deep
> Considered every creature.
>
> – *John Milton,* Paradise Lost

The term *text analytics* describes a set of linguistic, lexical, pattern recognition, extraction, tagging/structuring, 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 discover and present knowledge – facts, business rules, and relationships – that had been locked in textual form, impenetrable to automated processing.

Text analytics starts with document acquisition, either targeted retrieval of all material identified by a search or blanket intake of e-mail, Web pages, scientific papers, corporate reports, news articles, and the like. The next step is typically linguistic processing: determining sentence and phrase boundaries, stemming words, determining parts of speech. This step is followed by tagging and extraction of features – entities and their attributes, terms, concepts, sentiments, and relationships – with some form of term normalization and use of lexical analysis to provide frequency counts and the like. Use of taxonomies, lexicons and gazetteers, and machine-learning techniques facilitates this work.

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 discovered 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 and their contents. Information extraction, statistical analysis, visualization – none of these functions is present in typical search or content management offerings. Knowledge discovery – pattern recognition – via application of linguistic, statistical, and machine-learning techniques, and via data mining and visualization, is a key differentiator of text analytics from those latter technologies.

Because text analytics looks at document sets and identifies inter-document relationships, it supplies context that enables far greater relevance in 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-analytics 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.

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 European Text Analytics Summit and chaired the 2005 and 2006 summits in Boston and will chair the 2007 summit. His white paper for the 2005 summit, The Developing Text Mining Market, is available online at *altaplana.com/TheDevelopingTextMiningMarket.pdf*. His 2006 summit paper, Finding Value in Text Analytics, is posted at *altaplana.com/FindingValueInTextAnalytics.pdf* and his 2007 European Text Analytics Summit, European Text Analytics, from which Section 5 of this paper is excerpted, is posted at *altaplana.com/ EuropeanTextAnalytics.pdf*.

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

Text Analytics Summit 2007

The Text Analytics Summit 2007 (*www.textanalyticsnews.com/usa*), slated for June 12-13, 2007 in Boston, is a mindshare event for the leading developers, researchers, vendors, tech-savvy users, and newcomers to the text-analytics space.

This third annual Boston summit follows on the heels of a first European summit, held in Amsterdam in April 2007. Analyst Curt Monash wrote in *ComputerWorld* that "the [2005] Text Mining Summit ... was one of the best conferences I've been to in a long time." SPSS Vice President Olivier Jouve called the 2006 summit "the best conference I attended last year."

The European and North American Text Analytics Summits both provide an opportunity for researchers and vendors to identify promising applications, size up technical challenges, and connect with users eager to keep up with market developments. Text-analytics users and prospective users in any application or industry find an unmissable opportunity to learn from peers and understand the bottom-line impact of the latest deployments. Developers and marketers benefit from the opportunity to engage end users and technologist to better understand market requirements, technology developments, and product directions.

Alta Plana Corporation

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

