Text Technologies in the Mainstream

Text Analytics Solutions, Applications, and Trends

A white paper prepared for

Text Analytics Summit 2008

Boston, June 16-17, 2008
www.TextAnalyticsNews.com/4thannual08/

Seth Grimes

Alta Plana Corporation

Alta Plana

Table of Contents

This paper is dedicated to the memory of Joseph Weizenbaum (1923-2008), a pioneer in language-focused computing who advanced our understanding of the possibilities and limitations of information technology.

1 Introduction

Text technologies – text analytics, natural-language processing, and analytically rooted search – have entered the business mainstream, having evolved from research tools into business solutions. Appeal is stronger than ever for life-sciences researchers and intelligence analysts and other long-time users, yet in the last year, the center of attention has continued to shift in the direction of business applications such as customer relationship management (CRM), media and publishing, competitive intelligence, and financial analysis. Text technologies are allowing marketers to hear the Voice of the Customer (VOC) by enabling analyses of social media and contact-center interactions; they are helping revolutionize approaches to long-established practices in the legal and financial worlds.

The last year has seen further advances in integration of text technologies both with business intelligence and predictive-analytics tools and with line-of-business applications. Hosted and on-demand text analytics and adaptation to service-oriented architecture (SOA) – Text Analytics as a Service – have reduced barriers to adoption, creating flexibility for business focused end-users and lowering costs.

Marketplace indicators are uniformly positive, with booming sales at established firms, a steady pace of acquisitions of text-analytics companies by larger firms, and continued emergence of new and innovative products and services vendors.

This paper covers recent solution and technology developments that demonstrate the emergence of text analytics into the business mainstream. It profiles important applications in –

- CRM and marketing
- media and publishing
- electronic discovery and legal and financial compliance

– and considers key technical developments –

- addition of "unstructured data" to the BI and predictive analytics stacks
- realization of the potential of natural-language processing
- Text Analytics as a Service

– that both help us understand recent developments and signal directions for future development. This paper concludes with a market outlook for the coming year.

2 Focus on Business

Text analytics technology is increasingly accessible to business analysts and available for mainstream business needs – for enterprise feedback management, competitive intelligence, CRM, reputation management – integrated with line-of-business applications and conventional analytical tools.

2.1 Voice of the Customer

Voice of the Customer (VOC) is a time-tested business concept that has gained new life through the application of text analytics.

VOC researchers seek to understand the totality customer needs and opinions, whether explicitly stated or indirectly implied. They probe both individual views and collective, market thinking (which we might term voice of the market). Important information is no longer found only in corporate-sponsored and internally held sources. Text analytics creates the ability to discern and capture the voice of the customer from online media, such as blogs and forum postings; from e-mail, chat interactions, and contact-center dialogues; and from surveys and other mechanisms for collecting customer feedback, complementing traditional transactional sources.

As utilized in marketing and Customer Experience Management (CEM) programs, VOC builds on traditional transaction-based CRM to create a fuller picture of customer and market attitudes about an organization's products and services. According to Sid Banerjee, CEO of text-analytics firm Clarabridge, industries such as travel, hospitality, and retail "live and die on customer experience."

Marketers, account representatives, and product managers in customer support, marketing, and quality assurance use VOC findings to ensure enterprise responsiveness and competitiveness.

2.2 Media and Publishing

Text technologies are key to effectively using the diversity of electronic media and, for media companies, to ensuring that published information reaches the intended audiences. Search-engine keyword indexing and social tagging are not enough. Publishers and information consumers alike are adopting the semantic search, processing, and analysis capabilities delivered by text analytics.

Capabilities that have been available in specialized, narrowly focused tools for several years are now being embedded in mainstream and market-vertical applications. These capabilities include:

- automated topic identification
- clustering
- classification
- entity, concept, and fact extraction
- link and network analysis for discerned entities and concepts
- metadata extraction (for instance, of author, title, publication date, and other descriptive elements)
- sentiment extraction and analysis of attitudinal information
- visualization and interactive data exploration.

2.3 E-Discovery and Legal and Financial Compliance

U.S. Federal Rules of Civil Procedure (FRCP), Sarbanes-Oxley, and similar legal and financial-control mandates – in the U.S. and worldwide – have accelerated legal- and financial-sector adoption of text technologies.

These regulatory mandates require retention of huge volumes of electronic stored information (ESI): e-mail, instant-messaging traffic, telephone logs, transactional data from operational systems, depositions, sound and video recordings, and all manner of corporate communications, documents and records. Keyword search isn't adequate when it comes to discovery and investigatory functions. Text analytics has therefore become an essential tool in the legal and financial domains for conducting investigatory and forensic work that entails the discovery, collection, management, review, and delivery of pertinent information.

> Aaron Brown, program director, Content Discovery and Search, IBM Information Management Software, believes "legal discovery is at the front of the pack of promising new text-analytics applications along with a related cluster of use cases around compliance and legal control. Enterprises are looking to get out from under the crushing costs of traditional legal discovery and likewise to reduce the risks they face from compliance violations."[1]

The treatment of compliance-related materials is highly formalized due to the rigor of admissibility standards and professional procedures formulated in the course of hundreds of years of evolving legal, accounting, and auditing practices.

Consider *discovery*, the legal-sector process whereby parties to a lawsuit request and provide documents and information that may be pertinent in litigation. Discovery management has been transformed by the computerization of evidentiary materials that may be pertinent in litigation, whether that information originates in electronic form (e.g., e-mail) or is scanned from paper and possibly transformed via optical character recognition (OCR). The result is termed *e-discovery*.

According to e-discovery consultant Tom Lidbury of law firm Mayer Brown, "clients are adopting technology rapidly to manage all these processes." In 2006, Forrester estimated e-discovery technology spending would grow from $1.4 billion in that year to more than $4.8 billion in 2011. [2]

Text analytics can support legal and financial analyses – both in formalized processes such as e-discovery and in audits and investigations – by discerning important entities, relationships, and sentiments in textual sources, by extracting significant features, and by supporting classification and analysis of documents and extracted information. Workflow adaptation – shaping the tools to established legal- and financial-sector work practices – has made all the difference in the early adoption that is leading to mainstream use of text analytics in these sectors.

1 Seth Grimes, "Q&A: IBM's Aaron Brown on Text Analytics for Legal Compliance," *http://www.intelligententerprise.com/showArticle.jhtml?articleID=206901413*

2 Barry Murphy, "Believe It — eDiscovery Technology Spending to Top $4.8 Billion by 2011," Forrester Research, Inc., December 11, 2006

3 Expanding Scope

> Early adopters have deployed [text analytics] in conjunction with their data mining and BI solutions. In fact, the vast majority [of 2007 survey respondents] stated that they were extracting text and combining it with structured information for use with specific data mining tools or BI solutions. Companies... are considering other configurations as well.[3]
>
> – *Fern Halper, Hurwitz & Associates*

Text technologies extend established analytics applications by providing the ability to source information from documents and the Web. They simplify information access via natural-language (NL) interfaces, especially in the form of advanced search and NL data query. Service-oriented architecture and hosted and on-demand deployment options speed deployment time and lower entry costs by allowing business users to concentrate on business questions rather than on IT. The result is an expanded role for text analytics in mainstream computing.

3.1 "Unstructured Data," BI, and Predictive Analytics

Text analytics is an answer to the "unstructured data" challenge. The scope of the challenge is expressed in the truism that 80 percent of enterprise information originates and is locked in "unstructured" form. How complete can an organization's BI or predictive-analytics program be if it does not accommodate and exploit a majority of business-relevant information?

That problem has been recognized for decades. Consider the first recorded definition of BI, which appeared in an October 1958 *IBM Journal* article by H.P. Luhn, "A Business Intelligence System[4]," that describes a system that will:

> ...utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the "action points" in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points.

We see that the earliest BI focus was on text – on extraction, categorization, and classification – rather than on crunching numerical data! The explanation why BI refocused on numerical data is obvious: Organizations of all sizes harnessed computers to run their businesses, generating volumes of fielded transactional information that was (and is) directly tied to business operations. But now reporting and analysis of fielded, operational data has reached commodity status, and faced with the huge volumes of textual and other "unstructured" information available on the Internet, the imperative has shifted back to that original BI target: documents.

BI vendors have kept pace with the shifting market by integrating information extraction capabilities into their tools. Users can now analyze data drawn from "unstructured" sources using familiar BI methods and interfaces.

BI models are typically specified by business analysts. Predictive analytics, by

3 Fern Halper, Hurwitz & Associates, "Text Analytics: The Road to Understanding Your Company's Unstructured Information," 2007.

4 *http://www.research.ibm.com/journal/rd/024/ibmrd0204H.pdf*

contrast, applies models derived from statistical analysis and machine learning to discover patterns in data. Text analytics is a natural extension to established data mining and predictive analytics applications, just as it is for business intelligence.

It is increasingly common for data-mining users to seek to tap textual sources. Olivier Jouve, vice president of market strategy at predictive-analytics vendor SPSS, says the general trend is that 25-30 percent of his company's data mining customers also license text analytics capabilities. In Japan and in public-sector applications, the figure is over 50 percent.

3.2 Harnessing Natural Language

More than 40 years after Joseph Weizenbaum's ELIZA demonstrated an uncanny ability to mimic a patient-therapist conversation[5], natural-language processing (NLP) has come into its own as the basis for new interfaces and abilities including:

- question answering
- natural-language query
- enhanced search.

NLP allows systems to infer user intent in order to provide context and reduces the ambiguity that would otherwise degrade the value of search and query results. NLP-enabled search moves beyond reliance on keyword matches via an understanding of semantics, of previously hidden meaning. NLP enables systems to supply answers where previously they only returned documents.

While NLP and linguistics have long been the basis for automated text analysis and information extraction, broad availability of robust natural-language user interfaces is an innovation.

3.3 Text Analytics as a Service

Hosted/on-demand text analytics is the preferred service-delivery model for a rapidly growing segment of text-analytics users. This approach is a variant of the increasingly popular Software as a Service (SaaS) model. It comes in a number of flavors. With hosting, you contract another organization to install and maintain software for your use. With on-demand, you pay for only the services you use. Both styles of SaaS may be delivered via Web services interfaces or via more-traditional integration approaches, that is, where the client in effect outsources software automation of selected business processes.

Adoption of "as a service" text analytics, like other moves to SaaS models, is a response to efficiency, cost, and rapid-delivery business drivers. Users don't have to install and manage software, allowing a focus on business goals rather than on IT. They can get started quickly, relying on the provider's expertise, without incurring full software-licensing and training costs. SaaS adoption therefore not only represents a change in how software is used and paid for; it is a key enabler accelerating the adoption of text analytics by organizations that would otherwise lack the resources to license, design, and manage in-house solutions.

Fern Halper, an analyst with Hurwitz & Associates, said that a text-analytics survey conducted by her firm last year[2] asked if respondents would consider or have

5 Joseph Weizenbaum, "ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine," *http://i5.nyu.edu/~mm64/x52.9265/january1966.html*

implemented an "as a service" text analytics implementation. "A little more than half [of respondents] said yes. Twenty percent don't understand enough about the SaaS model and 12 percent replied 'don't know.' Only about 16 percent said no."

Some on-demand text-analytics users aren't even aware that they're using text analytics, for instance:

- users of online content search and analysis services and clients of information-services providers
- users of a spectrum of leading publishers and media outlets
- users of a spate of new entrants in the media-monitoring and reputation-management spaces.

These are essentially research services supporting all manner of information consumers, typically using custom-developed text-analytics and information-retrieval solutions. Here we see another promising direction for Text Analytics as a Service (TAaaS): Emerging TAaaS products will make it easier for solution providers to offer analytical capabilities, packaged with domain-related information and workflow, to customers whose primary concerns are business needs rather than technology.

4 Market Outlook

> The author estimates a 2007 worldwide market for text-analytics software licenses, support, and professional services of about $250 million with 25 percent annual growth through 2010.

The $250 million 2007 text-analytics market includes product vendors, technology suppliers who offer linguistic and natural-language technologies, industry vertical solutions, and an allotment of the portion of other-analytics revenue that is attributable to text analytics.

Growth during 2007 at new and established companies integrating text with BI was especially strong, and text-analytics companies focusing on particular industries such as life sciences and publishing also did very well, with sales expansion far above the sector average. Other companies focusing on technology – on toolkits and OEM licensing– did well in 2007, but growth in this segment, while reaching double digits, was below the sector average.

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 Consolidation

Last year, we were told to:

> 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.[6]

In the time since, Business Objects acquired Inxight (and was in turn bought by SAP), Reuters acquired ClearForest, SAS acquired Teragram, and Microsoft bought Fast Search & Transfer, FAST, an enterprise search vendor with significant text-analytics capabilities.

Two established text-BI suppliers, Attensity and Clarabridge, remain independent, as does Basis Technology, which earns significant revenues by licensing their linguistic and NLP technologies, competing with Inxight and Teragram. These vendors as well as established companies with text-analytics stacks such as Expert System, Lexalytics, Linguamatics, and TEMIS must be considered possible acquisition targets, as should vendors with attractive solutions targeted to legal-sector, financial, media and publishing, and VOC applications.

4.2 Emergence

In recent years, text-analytics entrants emerged most frequently via commercialization of academic or industrial research. In the last year, the bulk of new entrants appear to have been companies targeting particular verticals, such as the legal sector, or particular business functions, such as social-media analysis. This shift has been fueled by lowered barriers to market entry related to technology and is likely to establish itself as the rule: We will continue to see a preponderance of application-focused entrants to the text-analytics market.

6 Seth Grimes, "What's Next for Text," *http://altaplana.com/WhatsNextForText.pdf*.

5 Appendix: Technology Basics[7]

> And some certain significance lurks in all things, else all things are little worth, and the round world itself but an empty cipher.
>
> – *Herman Melville*, Moby Dick

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 annotate, extract, 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; data extracted to databases may be analyzed via BI, data mining, and visualization. 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 interdocument 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.

7 Adapted from Seth Grimes, "What's Next for Text," *http://altaplana.com/WhatsNextForText.pdf*.

Seth Grimes

Seth Grimes is a business intelligence, data warehousing, and decision systems expert, a consultant. He is Text Analytics channel expert for the *Business Intelligence Network*, contributing editor for *Intelligent Enterprise* magazine, where he focuses on business intelligence and advanced analytics, and a *Data Warehousing Institute* instructor. (Elements of this paper are drawn from articles of Seth's previously published in those outlets.)

Seth founded Washington DC-based Alta Plana Corporation in 1997 and consults on business intelligence and analytics strategy and implementation for clients in the U.S. and internationally.

Seth writes and speaks on data management and analysis systems, industry trends, and emerging analytical technologies. He is founding chair of the Text Analytics Summit series. His white papers for the 2005, 2006, and 2007 summits are available online at *altaplana.com/articles.html*.

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

Text Analytics Summit 2008

The Text Analytics Summit 2008 (*www.textanalyticsnews.com/4thannual08/*), slated for June 16-17, 2008 in Boston, is an essential event for the leading developers, researchers, vendors, tech-savvy users, and newcomers to the text-analytics space.

This year's is the fourth annual summit, latest in a conference series that has grown in size, scope, and impact with each successive year. The summit is designed to benefit practitioners, managers, and executives whole face the "unstructured data" challenge: the imperative to exploit valuable business information locked in previously inaccessible textual form.

The summit provides 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

