Alta Plana

SAP

From Text To Business Insight: *All About Enterprise Text Analytics*

Seth Grimes

Alta Plana Corporation

Sponsored by SAP

June 2011

Executive Summary

Text analytics automates the extraction of business value from Web sites and social media content, from e-mail, survey responses, and contact-center notes, from regulatory filings and corporate documents. These sources and others capture information from and about customers, prospects, products, companies, and competitors in human readable form. Examples range from restaurant and hotel likes and dislikes posted to online review sites to proprietary, confidential company or customer information disclosed in e-mail addressed to someone outside your company. There are dozens of scenarios where text analytics can help users both identify and exploit opportunity and detect and minimize risk.

Prospective users can choose from among a variety of tool and solution options. Users who are technically inclined may prefer software libraries or text data mining workbenches. Users in line-of-business departments will typically prefer a solution that adds analytics to the enterprise applications they use in their everyday work. The common point is the use of advanced linguistic and statistical algorithms to automate handling of business-critical information. **The differentiators, when it comes to enterprise adoption, are first, ability to interoperate seamlessly with enterprise operational applications and to extend enterprise business intelligence and analytics programs and second, provisions for the performance, reliability, scalability, security, and manageability required by competitive organizations.**

Enterprise-scale BI and analytics, for both text and for transactional and operational data collected in databases, are best built on a data-services foundation. This data-services foundation is comprised of a unified set of data, processes, standards, policies, and tools that ally to support flexible, agile, scalable, and secure systems development and deployment. **Data quality, profiling, and cleansing services – and data management, integration, and analysis facilities – all extended to handle social, online, and enterprise text, are essential foundational components.**

Ability to flexibly handle both scale and diversity – multiple languages, both *noisy* social sources and formal business and legal documents, factual information and subjective opinions, attitudes, and sentiment – is an additional *must* for enterprise text analytics.

Leading-edge organizations already exploit analytics-derived insights to drive operational and strategic decision making. **Ultimately, text analytics will form just one part of a larger analytics solution aimed at complete customer and business awareness, awareness not for its own sake but in furtherance of customer satisfaction, product and service quality, competitiveness, and profitability.**

Text Analytics and the Unstructured Data Challenge

Online, social, and corporate systems capture immense business value in textual form. Text-analytics solutions unlock that value. Enterprise text analytics takes solutions a step beyond, to enterprise scale, integrating with enterprise systems and business processes.

Text sources of interest include e-mail and text messages, blog and forum postings, survey responses and contact-center notes, warranty claims, and a wide variety of corporate documents. The information content is the raw *voice of the customer*, *employee*, or *patient*, *of the market* or *public*, depending on the medium and the relationship. **Business outcomes include better-targeted marketing and content delivery, higher customer satisfaction, quality improvement, competitive awareness, regulatory compliance, and brand-reputation security.** These outcomes relate directly to enterprise profitability, to risk management and the ability to identify and act on emerging opportunities.

Enterprise solutions

Enterprise IT solutions are different from solutions designed only for individual or departmental use**. Enterprise text-analytics solutions both handle a broad set of the narrowly defined tasks that more focused solutions cover – social-media monitoring and customer-survey analysis are examples – and further meet exacting reliability, manageability, and security standards.** They integrate, flexibly, with enterprise business processes and operational and analytical systems. They deliver needed performance and throughput when operating at enterprise scale.

Text analytics processes

Every application of text analytics involves **a few basic steps: 1) text acquisition; 2) preparation, processing, and storage; 3) analysis; 4) use of findings to further business goals; and 5) measurement of outcomes** with the aim of optimizing both business operations and analytical processes. Solution quality makes all the difference, in usability, usefulness, and ability to respond to business challenges.

The buttons and dials of usable technology are invisible when details can be safely hidden from end users but available when needed for when configuration and tuning will help users meet distinctive business needs. Usable and useful enterprise technology is not *siloed*, it doesn't create new islands of information. Instead, it introduces new data and capabilities to enterprise computing environments to create a whole greater than the sum of the parts. It helps solve business problems by **delivering insights aligned with business goals, insights that enable better decision making**.

Goals: Understanding and action

This paper seeks to help the reader understand the enterprise text analytics difference. It starts with an introduction to the technologies and business case and continues with a discussion of key enterprise considerations. It presents usages scenarios and concludes with a look at steps readers can take to build competitive value through enterprise text analytics: Basic implementation best practices that will guide the reader from understanding to action. The reader will come away with a better understanding of the *what*, *why*, and *how* of Enterprise Text Analytics.

Taming Text

Text sources contain information of immense business value. Our task is to discover, extract, and exploit insights, working not in isolation but rather by extending established and proven enterprise operational and analytical systems.

Reviews posted to public forum sites will tell you about perceived strengths and weaknesses of your company's (and competitors') products and services, down to the "feature" level (iPhone battery life, Verizon network speed in Los Angeles, the pastrami at Katz's Delicatessen in New York). You will find similar data, direct from customers and prospects, in survey responses, contact-center notes, and other forms of enterprise feedback, also in Facebook wall posts and Twitter updates and other *social* messages. These are important opinion sources. They often signal consumer *intent*, for instance, to make a purchase or cancel a subscription.

Information from social, online, and enterprise sources may be time-sensitive, localized, and highly subjective. Handled well, it is without equal in explaining the *Why* behind purchases, complaints, returns, and repeat business. The volume of messages however, leads to perceptions of *information overload*. How do we cope?

Dealing with information overload

Information overload is the notion that we are struggling to keep up with explosive growth in the volume of structured and unstructured data, data generated in the course of business and personal transactions, in human communications (social and news media, e-mail and messaging, documents of all forms) and by automated monitoring and recording (devices, log files, and audio and video).

Information overload is not a new concept. It dates at least to 1964, to Bertram Myron Gross's The Managing of Organizations: The Administrative Struggle[1]. Then, as now however, in Clay Shirky's words, "It's Not Information Overload. It's Filter Failure."[2] Filters reduce and channel the flood. Rework business processes and use automated software tools, including text-analytics solutions, to turn information into an asset rather than a distraction.

Information in text

Automated text technologies help us collect, select, manage, and make sense of just that data that can best serve as a source of business insights, filtering out spam and other noise. They extract, transform, and analyze information content that includes:

- **Descriptive values, also known as metadata.** Metadata may include author, creation date, language, and keywords. Metadata is often captured automatically, for instance by a Word processor or electronic camera.
- **Names of persons, organizations, geographic areas, products, and other entities**.
- **E-mail and postal addresses, dates, telephone numbers, account numbers, ticker symbols**, and other information indicated by patterns.
- **Facts, events, relationships, and sizes or measured amounts**: information that describes or interlinks entities.
- **Conceptual groupings of people, products, or brands**: teenagers, toys, and

[1] *http://bit.ly/h2mLf7*

[2] Cory Doctorow, "Clay Shirky on information overload versus filter failure": *http://boingboing.net/2010/01/31/clay-shirky-on-infor.html*

European auto makers, for example.

- **Structured data from data tables** embedded in Web pages and documents.
- **Topics and themes**: What a person would say a given text is about.
- **Sentiment – attitudes, emotions, mood, and opinions** – associated with potentially any of the features given above.

Text analysis can be challenging given complexities that range from misspellings to misinformation and from slang to sarcasm and given narrative, argument, and conversation that involves multiple voices and topics, over time, referred to via multiple names or labels that may include *anaphora* such as pronouns.

Leading solutions apply *natural language processing* technology that detects *named entities* and patterns that indicate other *features* of interest. They resolve parts of speech and further unravel phrase and sentence syntax in order to identify and extract facts and relationships. They transform text into data for further analysis.

Value beyond text

Names, facts, sentiment: Great stuff, important in themselves and as raw material for seeing connections, for generating business insights, relaying the raw *voice of the customer*, *employee*, or *patient*, *of the market* or *public*, and illuminating root causes behind customer actions and business decisions.

Additional business value lies outside any single text document or message, and even beyond any corpus (set of documents or messages). It is uncovered when you build text analytics into business operations and link text-sourced information to actual customer profiles and business transactions, to public relations or marketing campaign results, and to other records and real-world performance measures. Specifics depend on your *use cases*, on the business challenges you seek to solve.

Extraction and integration

Enterprise-class tools will match and marry qualitative, text-sourced information both to data captured in the course of operations and to market and reference data. You may choose to integrate at an application, record, or data-point level, or all three.

- Application integration builds text analysis into operational systems, often via an API (application programming interface). A local or remote annotation service marks up the *features* of interest in text passed via the API.
- Record linkage matches text (e-mail, online reviews) to customer profiles and transactions. It augments the information in a given message, for instance, associating past customer purchases and returns with service complaints or providing a transaction-derived estimate of customer lifetime value.

Given the anonymous nature of online postings, the precise identities needed for exact matches may not be available. Identify resolution remains a challenge. Data integration can nonetheless enrich analyses that focus on product and service features (rather than on individual customers), for example, by matching aggregate sentiment to marketing campaign results, sales figures, or share-price movement.

These are examples of *unified analysis of text and data*, worth further exploration.

Rounding out the picture, text analytics offers users the ability to automate content categorization, abstracting, and summarization – further, text analytics extracts *meaning* that fuels next-generation *semantic* search and data integration – all useful capabilities when it comes to advanced knowledge management and information access initiatives.

Text and Enterprise Analytics

Most enterprises have business intelligence initiatives in place, often complemented by predictive analytics, but those applications rarely take advantage of *unstructured* business content. Instead, analyses of information-rich document, e-mail, *social*, online, and other text sources are handled via siloed applications (if at all), at a departmental level, without any link to the operational data that fuels BI and predictive analytics. Worse, much of the software on the market for social-media and survey analysis offers only the most basic text and sentiment analysis capabilities, masked by slick but shallow dashboard interfaces.

Integration imperatives

***Unstructured* content and enterprise data holdings can and should be jointly handled, via data or application-level integration**. While you do need capable software to answer this imperative and you may need to re-examine business processes, the **complete customer/supplier/market/competitor awareness benefits** of integration can be very significant. **Beyond analytics, enterprises should embrace *information governance* practices for content – the *unstructured* analogue of enterprise data government and master data management (MDM) – as well as proactive customer *engagement* practices** that reflect new ways of operating in always-on online and *social* environments, essential practices for *social CRM* done right.

Integration points

We stated earlier that every application of text analytics involves a few basic steps: 1) text acquisition; 2) preparation, processing, and storage; 3) analysis; 4) use of findings to further business goals; and 5) measurement of outcomes with the aim of optimizing both business operations and analytical processes. Each of these steps can be an integration point, for unified information processing, management, analysis, and presentation.

- **Text acquisition** may involve identifying and retrieving material from online or *social* sources, via crawling, indexing, and search. Alternatively, it may involve bringing in reports, warranty and insurance claims, e-mail, contact-center notes, and other content from corporate systems and repositories, via some form of adapter or interface and again, possibly involving search. In either case, it is important to design-in collection of identifiers and other metadata that can later be used for record matching.
- **Data profiling and cleansing** are key data-preparation steps, essential to ensure data quality, whether dealing with text or with structured databases. Analogous concepts apply in the two realms: Controlled vocabularies = Master data; Taxonomies and facets (categorization schemes) correspond to Data dimensions; Metadata is metadata: Values that describe data/text provenance, type, authorship, use, and other attributes. In both domains, we have a concept of reference data and dictionaries that standardize terminology and value sets.

  In the name of consistency and efficiency, and also to facilitate unified analysis, coordinate these steps across text and structured data, and further, in work with information of either type, align information governance, information access, retention, inventory and audit, security and privacy, and other policies.

- It is broadly understood and accepted that bringing to bear **multiple analytical**

methods, tapping multiple sources – whichever methods and sources are relevant to the business challenge at hand – **creates analytical *lift***, insights that are more accurate and usable than can be obtained via the use of any single method or source. Plan for unified search, information access, analysis, and presentation.

- Mainstream BI and analytics are all about numbers and indicators (some of which may be text sourced) whether delivered via dashboards, reports, visualizations, or pivot interfaces. Text sources provide **qualitative data that not only explains the numbers; they contain distinctive *signals* not present in numbers-only results**. Text is at the center of an emerging class of search-based applications, which are often line-of-business focused and business-process embedded, drawing from both text and structured data sources.
- And of course, every business initiative should be justified via in terms of contribution to business goals. We should provide quantified measures of text analytics' contribution where feasible – a bit complicated given, for instance, the indirect linkage of *social* mentions to marketing-campaign, customer relationship, and sales results – and we should recognize **text analytics' contribution to achieving non-direct ROI, to boosting customer satisfaction, response time, quality early warning, community building, and other enterprise goals that are not measured by traditional means**.

A Data Services Foundation

Enterprise-scale computing applications are best built on a data-services foundation, on a unified set of data, processes, standards, policies, and tools that ally to support flexible, agile, scalable, and secure systems development and deployment. With designed-in scalability (to handle new users and users), extensibility (to new information sources and types), unified data services deliver efficiency and manageability that cannot be matched with a hodge-podge of disjoint systems. Business benefits include reliability, cost control, and a level of capabilities that accelerate speed to insight and ROI.

The foundation will unite data quality, metadata management, data transformation and integration, and data management capabilities to provide a services framework for what is commonly known as *trusted data*, of course extended to text.

Data services provide access to enterprise information on-demand, as needed, without regard for data type or source, through application programming interfaces (APIs). Behind the scenes, back-end systems record provenance, cleansing and transformation steps, and utilization to support compliance and service-level monitoring, audits, and steps to optimize management and resource utilization. They apply uniform data access, privacy, and security policies.

No limits

Uniformity and unification are good, but they shouldn't be a straitjacket that limits ability to respond to special challenges best met with specialized data, algorithms, or processes. Given the variability of human language – differences across cultures, channels, and business domains – enterprise text analytics must accommodate needs that may include:

- **Multi-lingual and cross-lingual text**: To state the obvious, not everyone speaks (and writes in) English or even uses a Latin alphabet. Translation won't get you far, especially when you're dealing with slang, jargon, and sentiment in *social*

sources.

- **Domain adaptation**: A Facebook wall post, an SEC filing, and an intelligence report may all be written in English, but the language usage – the vocabulary and syntax and therefore the interpretation – and information content will very significantly differ. Context sensitivity boosts interpretation accuracy.
- **Beyond-polarity sentiment**: Many prominent sentiment-analysis tools force classification into positive/negative/neutral polarity bins, and too often they score sentiment at a message or document level and don't resolve feeling about particular concepts or feature. Analysis should be aligned with business needs; for instance, wouldn't a happy/angry/sad emotion classification be advantageous in a customer support applications? Opinion mining is another useful form of beyond-polarity sentiment analysis.
- **Signal detection**: Many dedicated social-media analytics tools are infamous for their shallow focus on counting mentions and projecting trends, leaving it to the analyst to guess at a link to actual business outcomes. For true Social CRM – the ability to personalize and profit from engagement across touchpoints – we need the ability to mine signals: Intent to buy a product, recommend a supplier, cancel a service, or, for that matter, commit a crime, while there is still time for intercession. Analysis tools must, clearly, adapt to the user's needs.
- **High-velocity analysis**: Machines operate 24/7 with ability to automate processing of huge volumes of information. Reaction time is key to ability to profit, whether you application is automated reputation management, financial-markets trading, or counterterrorism.

  High-velocity analysis has been the domain of complex event processing (CEP) technology, which typically processes, transforms, and analyzes streams of numerical data such as share price quotes and offers and trades. "No limits" means adapting CEP capabilities to text, capabilities that include event-driven analytics, real-time action in response to signals detected in source.

Ability to flexibly handle both scale and diversity is a key attribute of enterprise text analytics.

Usage Scenarios

Usage scenarios help prospective users envision how they might beneficially adopt a new technology at their own organizations. Whatever the scenario, however, in the words of Philip Russom of the Data Warehousing Institute, "Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before."[3] We examine three scenarios.

Quality early-warning

Awareness of quality issues may be slow to emerge if you rely exclusively on traditional feedback channels, on indicators such as product returns. Text analytics lets you bring new data into play and it lets you better exploit existing resources, creating new quality early-warning possibilities. Online social platforms are a fantastic source of quality insights, especially for consumer goods and services where customers are quick to post experiences and opinions to social sites. Quality early-warning starts with monitoring and measurement – start by looking for brand mentions associated with quality-related terms – but it requires sophisticated text analytics to correctly extract relevant details, facts, events, and opinions and, where possible, match extracted information to inquiry, purchase, service request, and other transactional records. Further, text analytics allows large-scale mining of warranty claims and internal reports in a search for patterns that not only indicate issues, but that also hint at root causes.

Risk management, fraud, and compliance

Search is typically reactive, a tool for after-the-fact, forensic investigation of suspected fraud and compliance incidents. To better manage risk, we want to move fraud detection and compliance monitoring into real time, via operationally embedded text data mining.

Text data mining here means information extraction, pattern detection, scenario modeling, and model deployment for active surveillance. We mine for words and phrases that suggest fraudulent health-insurance claims and theft reports; disclosure of sensitive or propriety corporate information; and other irregularities. Models that extend to customer profiles and reference data sources will be especially well equipped to detect *outliers*, suggestive, anomalous behaviors. Of course, modeling and detection are only part of a risk/fraud/compliance solution, which also requires a rules engine to automate processing and response, whether a simple alert of a more complex action. In-memory analytical technology can help speed execution.

Retail demand forecasting

Demand models built around inventory, orders, and past sales fail to account for fast-emerging disruptions that may stem from competitive developments or market externalities. Round out demand models with news, opinions, and *intent signals* harvested from online and social media, via automated text analytics. You will boost model accuracy, reliability, and timeliness.

Text and network analysis can detect and quantify online buzz as well as the forms of content and the influencer profiles that are capable of moving markets, allowing organizations the possibility of shaping, and not just reacting to, demand.

[3] BI Search and Text Analytics, *http://tdwi.org/~/media/TDWI/TDWI/Research/BPR/2007/TDWI_BPR_2Q07_BISTA%20pdf.ashx*

Implementation

The saying "A journey of a thousand miles begins with a single step," attributed to the Chinese philosopher Lao Tzu, aptly describes a best-practices approach to text-analytics implementation: Deliberate and progressive, with later steps building on earlier. The journey's not arduous, but it will involve planning, a focus on goals, and care to find the right path. Benefits are compelling and there's no good reason not to just take that first step.

For new adopters, it is often wise to begin with a modest initial effort and build out from there. You can start small, with a focus on particular source materials and business needs. Source materials are readily available and have clear value, whether social and media postings about you company and products, internal e-mail and documents, or survey responses and contact center dialog. You will quickly learn by doing and will discover the approach that will help you meet your organization's needs.

First steps

Step one is to define those business needs. Next, form an idea of the insights that will help you meet those needs and which source materials can be mined to gain those insights. Then sketch the analysis processes that will transform inputs into insights and the operational processes that will exploit those insights to drive decisions. Also figure out how you will measure outcomes and evaluate results. Lastly, identify software (or Web services) that can help you do the job and set up a pilot, proof-of-concept implementation.

These eight initial steps will get you underway at low risk and expense and will create a framework you can build out into a broader, deeper solution.

Enterprise solution design

Your organization already exploits analytics-derived insights to drive operational decision making. Ultimately, text analytics will form just one part of a larger solution aimed at complete customer and business awareness, awareness not for its own sake but in furtherance of customer satisfaction, product and service quality, competitiveness, and profitability.

Don't needlessly create parallel, duplicative, disconnected decision systems just for text. Instead, aim to fold text into existing BI/analytics efforts. Because you already have systems in place, text-analytics capabilities provided by current suppliers, or by their technology and solutions partners, will often be quickest and simplest to implement. They will have a degree of integration with already-in-place systems and you will be able to exploit existing business arrangements.

There are many alternatives available from service and software providers. If you do go outside your current supplier network, do prioritize integration (data, systems, and business process) and enterprise experience and scalability as evaluation criteria. Learn from experiences at organizations comparable to yours that face business and technical challenges similar to your organization's.

Directions: Enterprise Futures

Often we can paint the future in broad strokes even while particulars are cloudy, less certain. The path toward computing-communications convergence was clear from the earliest days of the Web, although no one in 1993, the year the NCSA Mosaic Web browser launched, could have foretold the form and details of the cloud-backed, mobile, social, text-rich computing that is the current decade's obsession.

Steep growth in the volume of data created and stored will continue unabated. Personal devices, sensors, and social and online platforms will continue to generate structured data records and unstructured text, images, audio, and video. We already capture this Big Data for after-the-fact retrieval and analysis. We will increasingly seek to analyze it in-flight, as it is produced, in order to respond to opportunities and threats in real-time, as they emerge. *Big* and *fast* – and add a concept of *wide*, of correlated data and events from disparate, distributed sources – sum up to *complex*, to **an era of Complex Data, supporting personalized information delivery that is intent-driven, task-oriented, and contextually aware.**

Complex Data, semantics, and situational intelligence

Complex Data has many sources. In business contexts, elements include customer profiles, transaction records, and reference data. Add to the mix online, social, and machine data: locations; text-extracted facts and opinions; search and query logs; and surveillance, clickstream, and tracking data – data derived from enterprise, machine, and inter-personal actions and interactions, data that can be mined for subjective insights and as raw material for behavioral and psychometric modelling.

We interlink datasets, records, and individual facts and values to create a composite that marketers characterize as a *360-degree view* of the customer, market, or other subject. To automate link creation – to generate insights from multi-sourced data – systems require *semantics*, a term that is often equated with, but that is more than, *meaning*. **Semantics forms the basis for flexible knowledge structuring, for dynamic data integration, and for automated inference to meet situational needs. The technology that discerns semantic meaning in free-form sources is... text analytics.**

Situational applications take into account circumstances and intent: Where and who we are and what we hope to accomplish. They use context to select and disambiguate sources, generate goal- and outcome-focused insights, and present information in appropriate, usable forms. An augmented-reality mobile app that overlays a street scene with restaurants and retail outlets is situational. So is an automated securities trade triggered by a combination of share-price movement and news-harvested sentiment scores, and so is Web advertising chosen on the basis of page context, site-visitor profile, and past Web-site visits.

Erasing boundaries

Situationally focused online-social-mobile computing redefines and even erases boundaries between corporate and personal systems. We tell our Facebook about our everyday lives, we describe our restaurant experiences on Yelp, and we tweet our politics, customer-service problems, and the music we listen to... and seek to bring a similar, collaborative experience to our work lives. Analytical systems that interpret and make business sense of human communications helps organizations cross boundaries and redefine relationships. Enterprise text analytics plays a role that will only grow in importance.

About

Seth Grimes

White paper author Seth Grimes is an information technology analyst and analytics strategy consultant. He is contributing editor at TechWeb's *InformationWeek*, founding chair of the *Text Analytics Summit* and the *Sentiment Analysis Symposium*, an instructor for *The Data Warehousing Institute (TDWI)*, and text analytics channel expert at the *Business Intelligence Network*.

Seth has worked with database, BI, and decision-support applications and users for over 25 years. He founded Washington DC-based Alta Plana Corporation in 1997. He consults, writes, and speaks on information-systems strategy, data management and analysis systems, industry trends, and emerging analytical technologies.

Seth can be reached at *grimes@altaplana.com*, +1 301-270-0795. Follow him on Twitter at *@sethgrimes*.

SAP

As market leader in enterprise application software, SAP (NYSE: SAP) helps companies of all sizes and industries run better. Founded in 1972, SAP (which stands for "Systems, Applications, and Products in Data Processing") has a rich history of innovation and growth as a true industry leader. Today, SAP has sales and development locations in more than 50 countries worldwide. SAP applications and services enable more than 109,000 customers worldwide to operate profitably, adapt continuously, and grow sustainably.

Visit *www.sap.com*.