

# Text Analytics in the BI Ecosystem

Seth Grimes

*Alta Plana*

*Sponsored by*

**SYBASE®**

## Introduction

**Text analytics transforms text into data for BI, data mining, predictive analytics, and smarter search.** Solutions find meaning and business value in text.

### Unstructured sources

**Text analytics is a semantic technology that complements conventional business intelligence.** Where BI has long focused on numerical data, text analytics extends BI's reporting, analysis, and visualization capabilities to the estimated 80% of business-relevant information found in text and other "unstructured" forms. **Any written (or spoken) material can be analyzed;** sources may include:

- E-mail and text messages;
- Web pages, blogs, forums, and other social and news media;
- Contact-center notes, survey responses; warranty and insurance claims;
- Corporate reports and filings; and
- Legal documents and scientific literature.

We automate analyses given the computer's ability to work with large document volumes, fast, in multiple languages and in diverse file formats and forms ranging from terse and noisy (e.g., text messages) to formal (scientific papers).

### Deep information content

Text often captures straight-forward facts – variables and values – albeit embedded in *natural language* rather than stored in database tables. **The unique value in text, however, is information content that goes beyond facts – themes and topics, events and relationships, opinions and emotions – information content that explains the why behind the what and when recorded in transactional records.** Text analytics mines this content to enhance search, discern patterns, and reveal intent, relationships, and root causes.

In the words of Philip Russom, senior manager of research and services at the Data Warehousing Institute, **"Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before."**

### Business value

Text analytics applications date to the 1990s, in intelligence and the life sciences. In the years since, text analytics has had a significant impact for purposes ranging from *voice of the customer* applications to media and publishing and semantic search. **These successes have created a \$425 million global software and support market in 2009, representing billions of dollars of business value.**

**Organizations looking to get started with text have options that range from hosted and as-a-service to dedicated text-processing applications.** Integration with the enterprise data warehouse and with corporate BI and data-analysis systems is another, attractive option: it is a given that combining data from multiple sources and applying multiple analytical methods provides lift, the ability to see more than can be discerned with any one source or method.

### Text analytics: The way forward

This paper elaborates on business benefits and technical challenges related to taming text. It introduces text technologies and surveys noteworthy business applications. And it offers a roadmap to guide you in bringing text into your organization's BI environment to support more accurate and timely business decision making. **Text is BI's next important frontier. This paper explains why and how.**

## Taming Text

An estimated 80% of enterprise-relevant information resides in text and other “unstructured” forms. These materials are created by people, for people. They communicate facts, opinions, and other information, for personal and business purposes. When we talk, write, fill in forms, or create pictures, audio, or video, we generate information meant for others to hear, read, process, and understand.

Our thoughts and needs are captured in text when we write e-mail and text messages, business documents and reports, news and blog articles, product descriptions and advertisements, and scientific papers and business reports; when we fill out surveys and insurance or warranty claims; and also when spoken words are transcribed and when we label and describe images and video. Text captures deep and broad business value in ways that numbers and record-based computing systems, designed to process transactions, can not. Text contains factual, qualitative, and subjective information not present in conventional data systems and analytical data warehouses.

### The “Unstructured Data” Challenge

Text may contain easily recognized data, even data tables, as well as qualitative information that explains the numbers, whether business transactions captured in a database system or the results of a marketing or publicity campaign or the reason a customer returned a product or filed a warranty claim.

“Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically.”

– Prof. Marti A. Hearst, “Untangling Text Data Mining,” 1999

Automating access to this information is the *unstructured data challenge*. Automation is essential. Computer software runs 24/7. Systems ingest, process, and store very large volumes of data, work across languages and business domains, and deliver results for a wide variety of business applications. Responding to the text challenge involves –

1. Finding the right, business-relevant information.
2. Using inherent structure – the frequency and distribution of words, clues from syntactic elements such as parts of speech, and language patterns – to achieve the following:
  - Infer meaning.
  - Discern relevant information content.
  - Structure content for machine use.
3. Applying analytical methods to generate insight.
4. Presenting usable, actionable findings to support business decision-making.

### The role of text analytics

Text analytics responds to the *unstructured data challenge*. It automates the process of taming text.

People use text – Web pages, e-mail, social media, reports and documents, enterprise feedback – in diverse ways. We can reduce those ways to four high-level categories. What do we do with text? We –

- **Publish, manage, and archive** text including user-generated content.
- **Index and search** every type of document and message imaginable.
- **Mine and extract** information content.
- **Categorize and classify** according to metadata and content.

Text analytics generates descriptive and semantic *metadata* that captures essentials

such as document or message author, date, title, keywords, and description. Metadata may include extracted topics and themes. It enhances our ability to deliver **intelligent content**. Text analytics also creates accurate document summaries, enables high-quality machine translation, and enriches content via semantic annotations.

Text analytics enables search on concepts, driven by natural-language queries rather than (just) keywords, with improved relevance and usability of results delivered via clustering, faceted navigation, links for supplemental information and related queries, and other capabilities that add up to **intelligent search**.

Text analytics is the key to **information extraction** starting with *named entities* such as people, organizations, geographic places, and stock ticker symbols and pattern-based values such as phone numbers, dates, and e-mail addresses. Beyond entities, text analytics extracts concepts (Ford, Toyota, Fiat, etc. are “vehicle manufacturers”); topics and themes; facts, relationships, and events (“Sandy Koufax of the Los Angeles Dodgers pitched a perfect game against the Chicago Cubs on September 9, 1965.”); and opinion and sentiment (“Highly rate this hotel - from minute you walk through door you start feeling very spoiled. Superb room, excellent service.”)

And text analytics – text data mining – is instrumental for **knowledge discovery** in unstructured information sources, in natural language.

### A history of sense-making

Language is a tough nut for machines to crack, so how does text analytics work? Consider two very similar sentences with very different senses of identical words!

Time *flies like* an arrow. Fruit *flies like* a banana. (Groucho Marx)

Text analytics automates sense-making, and not just of quips, replicating what you and I – and researchers, writers, and scholars – have been doing for millennia. The technology has its roots in 1950s work done by researchers including IBM’s Hans Peter Luhn, who in 1958 defined *business intelligence* as automated analysis of text. In an IBM Journal paper, Luhn wrote,

*Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance.*

Statistical significance points us to key topics and themes. It is the basis of keyword search and search-engine optimization. But Luhn recognized even then, in 1958,

*This rather unsophisticated argument on ‘significance’ avoids such linguistic implications as grammar and syntax... No attention is paid to the logical and semantic relationships the author has established.*

Over the years, researchers have developed techniques to model text, including those elusive semantic relationships. Some map documents into mathematical *vector spaces* to facilitate clustering, classification, and similarity computations. Others rely on *lexicons* and *thesauri* to identify terms and synonyms and on word *morphology* rules for tasks such as *stemming* and *lemmatization* (*go, goes, went, and to go* have the same root) and the identification of parts of speech: the subject, verb and object (and more) that encode facts in human languages. Additional tools such as *semantic networks* help in disambiguation, the use of contextual clues to help determine if “ford” in a given phrase is a vehicle manufacturer, an actor, a president, or a place to cross a river.

Yet while technology is important, what counts most is results, business value, as we shall see in the sections that follow.

## A Focus on Applications

### Paths to insight

Text analytics excels in delivering business value, not only or even primarily in conventional ROI terms, but also in measurable forms relating to–

- Improved customer satisfaction and retention.
- Ability to identify and act on opportunities in real-time, as they emerge.
- New insights into fraud and risk, in domains ranging from on-line commerce to financial markets and even counter-terrorism.
- Faster and more accurate processing of inquiries, claims, requests, and casework.
- Ability to uncover root causes of product and service quality issues.

*“Text Analytics is exciting technology, opening up new applications and approaches to solving information needs and supporting decision making for an improved customer experience.”*

– Michael House, Maritz Research, Division Vice President

Users realize benefits in analyses that currently focus primarily on text – customer-experience management, semantic search, next-generation media and publishing, and (most) social-media analytics – and, especially, in analyses that integrate text into established BI and data mining initiatives. It is well understood that multiple methods, applied to multiple data sources, can generate analytical lift, the ability to see and understand more than you will working with any single method or source.

We will examine applications associated with three critical business needs.

### E-discovery and compliance

*E-discovery* rules govern the production, in the event of litigation, of legally relevant electronic documents and other evidentiary materials. The process is well-defined in legal practice; rules are exacting and penalties for non-conformance are stiff.

Beyond e-discovery preparedness, organizations must comply with regulatory mandates and also protect themselves against fraud, risk, and disclosure of trade secrets and other proprietary information. These needs dictate that organizations –

- Monitor corporate communications
- Manage electronic stored information (ESI)
- Provide means of searching and analyzing ESI content

– to support e-discovery production, ensure compliance, and detect and respond to emerging threats.

Search is key, and beyond basic search the ability to discern often-complex patterns – concepts, topics, and themes, and interactions over time that may involve multiple parties – in very large volumes of diverse material that include e-mail, text messages and chat, voice recordings and transcripts, and news and social-media postings. Text analytics strongly contributes to effective e-discovery and compliance.

Consider only solutions that collect, index, and support analysis of ESI in an accessible repository, in accordance with mandated procedures and proven compliance best practices, with enterprise-class management and search capabilities.

### Customer experience

Customer experience management (CEM) – a step beyond customer relationship management (CRM) – is a fantastic use case for text analytics. CEM draws insights from *Voice of the Customer* sources to understand experiences, motivations. Sources include enterprise feedback and customer interactions –

- E-mail, contact-center notes, and survey responses
  - On-line forum and blog postings and other social media
- similarly useful for next-generation marketing efforts, mined and analyzed via text technologies, which allow organizations to –
- Address customer product and service issues
  - Improve quality and refine product and service design and delivery
  - Engage on-line to manage brand image and reputation, to respond rapidly to emerging opportunities and competitive threats

*“We've uncovered concepts and relationships in text that would be too costly – or even impossible – to detect by any other methods. We can now combine multiple data sources to evaluate customer expectations and improve customer satisfaction by employing more one-to-one customer contact and pre-emptively resolving customer complaints to keep our retention rates high.”*

– Federico Cesconi, Cablecom, head of customer insight and retention

The ability to link customer feedback to individual profiles and transactions creates even greater possibilities. If you can link qualitative and quantitative customer information, perhaps by choosing a single data-analysis platform that manages and unifies analysis of text and conventional data, you can –

- Analyze customer perceptions according to demographic variables
- Assess customer value in order to determine appropriate problem resolution steps
- Trace problems to root causes, for instance, by identifying the staff who handled a problematic customer interaction
- Model indicators for churn likelihood, cross-sell and up-sell opportunities, etc.
- Anticipate and forestall issues

### **Market and competitive intelligence**

Market and competitive intelligence aggregate the experiences and opinions of individual customers and prospects – an organization's own as well as competitors' – to capture the *Voice of the Market*. Sources are the same as for customer-experience analyses: enterprise feedback and customer interactions. Add to those sources news and social media, mined for competitive information regarding pricing, promotions, positioning, and strategy. The goal is to understand, at tactical and strategic levels, opportunities, threats, and trends.

*Social-network analysis* is an extremely interesting application of text technologies coupled with data mining. Organizations wish to track, map, and predict the diffusion of messages across networks, topics that are not only *tweeted*, blogged, and posted but also *re-tweeted*, forwarded, and commented on. Who are the influencers, the thought leaders and connectors? What corporate messaging will they pick up on? How widely and how quickly will their views spread and where?

Network analysis helps in the design of marketing campaigns and in the formulation and placement of advertising and publicity. Other analyses can aid an organization in better categorization and positioning of product and service offerings for on-site search and faceted navigation, to enrich content delivery, and to optimize channel use. Possibilities are limitless, but how to get started?

## Getting Started with Text Analytics

*In a 2009 survey, “all of the companies that had deployed text analytics stated that the implementations either met or exceeded their expectations. And close to 60% stated that text analytics had actually exceeded expectations.”*

– Fern Halper, Hurwitz & Associates

Users have a wealth of text-analytics options. There are possibilities for every business need and situation, from single users to departmental and enterprise scale, for a variety of industries and business functions, for text that ranges from informal, voice of the customer sources such as Twitter, blogs, survey responses, and e-mail to formal materials such as corporate filings, scientific papers, and legal documents.

### The solutions marketplace

Text-analytics solution choices include:

- Pure-play software, hosted, and service options that focus solely on text.
- Data-mining workbenches extended to support text.
- Tools and services for software developers.
- Line-of-business applications that include text-analytics functions.
- Enterprise data-management platforms, including options that bring text into the database management system (DBMS).

Craft your evaluation and selection criteria to reflect business goals, user capabilities, information types and volumes, and, of course, budget and IT-resource availability. Do follow a best-practices approach to getting started that will ensure success.

### The analytical DBMS advantage

Many organizations will benefit by managing text within the DBMS, alongside data stored in conventional, relational data tables. (Sybase IQ 15.2 is the first of the leading analytical database systems to build in text search and analysis.) Unified storage of text and data offers several advantages over other approaches:

- DBMS security and reliability are extended to textual documents.
- A single query can reach both text and data.
- In-database analytics, offered by leading analytical DBMS vendors, is extended to text. Embedding and executing analytical functions in the DBMS rather than external programs has benefits that include a) faster performance and reduced security vulnerability due to elimination of data movement between the database and application programs, b) elimination of redundant analytics programming given centralized management of analytical routines, and c) simplified application development: programmers access analytics via simple SQL queries.

Of course, actual capabilities will vary by DBMS. Not every DBMS meets enterprise scalability and reliability needs, and not every DBMS vendor has significant experience with enterprise customers across major industries.

Scalability, robustness, and experience concerns apply regardless of the technical approach, whether an organization opts to store and analyze text in-database or not. For this reason, both technical and business factors are part of a best practices approach to solution design and vendor evaluation and selection.

### A best practices approach

*Best practices* draw from theory and experience to guide organizations in embracing new, innovative information technologies such as text analytics. Three basic getting-started steps will apply. Adopters need to *assess*, *address*, and then *implement*.

1. First *assess* in order to understand:

- Business drivers and goals.
  - Ways your use of text analytics will extend existing BI and analytics efforts.
  - Key stakeholders, systems, and business processes.
  - Available technical resources, information sources, and budget.
  - Relevant experiences at your organization and at organizations like yours.
2. Evaluate options and approaches that *address* needs, challenges, and concerns:
    - Vendor solutions, capabilities, experience, and business reliability.
    - Performance and business requirements and ROI measurement.
    - Analytical methods and approaches (installed, hosted/SaaS, packaged solution, or database-integrated) in light of goals, sources, and work practices.
  3. Plan, *implement*, test, measure, and refine... and then repeat to extend your solution.
    - Start with basics such as search; design for scalability and extensibility.
    - Go for clear early wins to gain support.
    - Build out applications, capacity, BI/analytics integration, and user base.

It is important to choose a solution provider that can keep pace with evolving requirements. The right choice, sensibly implemented, will help you tame text to extend BI and analytics capabilities for better enterprise decision making.

## About

### Seth Grimes

White paper author Seth Grimes is an information technology analyst and analytics strategy consultant. He is contributing editor at *Intelligent Enterprise* magazine, founding chair of the *Text Analytics Summit*, 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](mailto:grimes@altaplana.com), +1 301-270-0795.

### Sybase

For 25 years, Sybase has been a leader in developing and expanding innovative database technology. Since its founding in a Berkeley, Calif., home in 1984, Sybase has earned the trust of many of the world's leading companies for its ability to manage information and deliver unsurpassed levels of data reliability and security. Today, Sybase is the largest enterprise software and services company exclusively focused on managing and mobilizing information. With its global solutions, enterprises can extend their information securely and make it useful for people anywhere using any device.

Sybase

One Sybase Drive

Dublin, CA 94568

Phone: +1 925-236-5000

Visit [www.sybase.com](http://www.sybase.com).