

UNCOVERING INSIGHT HIDDEN IN TEXT

How Unstructured Data Analysis Is Helping IT Deliver Knowledge

CONTENTS

- 1 Hidden Treasure Inside the Enterprise
- 3 Challenges Delivering More Insight to Business Units
- 4 Brief History of Text Analysis
- 6 Gaining New Insight with Unstructured Data Analysis
- 7 The Benefits of Unstructured Data Analysis
- 8 How Unstructured Data Analysis Works
- 9 What to Look For in an Unstructured Data Analysis Solution
- 10 The Business Objects Solution
- 11 About Business Objects

As the shepherds of corporate data, IT provides business units access to critical information that molds the future. With insight from business systems, decisions are made and courses are set. However, what if only half the picture is being examined?

With business units demanding more access to corporate data, IT is pressured to deliver. Yet current business intelligence (BI) and data warehouse systems only examine specific quantifiable data. Equally important (yet hard to analyze) qualitative data is quietly passed by. This treasure chest of information includes comments on forms, email communication, blog posts, and other text-based data. When companies overlook unstructured information, business decisions are made with limited knowledge.

How can IT provide departments insight hidden in text? How can your BI and data warehouse investments leverage unstructured data?

Designed to overcome these obstacles, unstructured data analysis solutions are providing business units new levels of insight without taxing IT. These applications intelligently extract knowledge from text and present understandable results to BI and data warehouse systems. Now, businesses can holistically examine the entire set of available knowledge on a topic, adding greater confidence to their decisions.

This white paper from Business Objects, an SAP company, explores the growth trend of unstructured data and examines the many benefits of unstructured data analysis.

More than half the data in an enterprise lies in untapped textual content.

HIDDEN TREASURE INSIDE THE ENTERPRISE

Extracting knowledge hidden in text will play a key role for businesses in the near future. According to Merrill Lynch, as much as 85% of corporate data is hidden in hard-to-access documents, such as “e-mails, memos, notes from call-centers and support operations, news, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations, and Web pages.”¹ A more conservative estimate from the Data Warehousing Institute (TDWI) finds unstructured and semi-structured data amounts to a smaller yet significant 53% of all enterprise data.²

“Regardless of how the numbers add up, we all know that the average user organization has a mass of textual information that business intelligence, data warehouse technologies, and business processes are ignoring,” says senior TDWI researcher Philip Russom.³

¹Blumberg, R. & Atre, S. (February 2003). DM Review. The problem with unstructured data.

²Swoyer, S. (September 5, 2007). Enterprise Systems. Unstructured data: Attacking a myth.

³Ibid.

The mass of knowledge stored in unstructured data will not disappear. IDC predicts, “Most of the digital universe will remain unstructured – meaning tools and techniques will be required to add structure to this content to improve search, discovery, management, security, and storage.”⁴

An IDC report explains that the analysis of text-based data will play a key role in regulated industries and will help many businesses cut call-center costs and increase sales.⁵

According to a TDWI survey, unstructured data sources will play a growing role in data warehouses in the coming years.⁶ The study cites information in email, word processing files, web pages, RSS feeds, and instant messages as hot targets for unstructured data analysis.

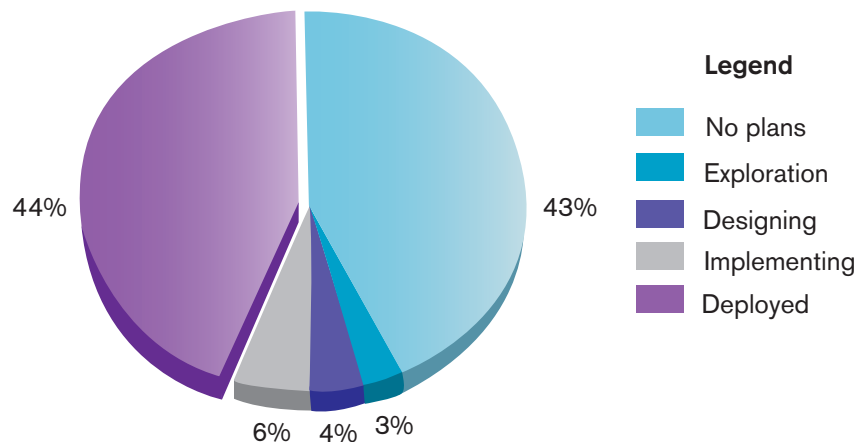


Figure 1: Shows 56% of Enterprises Employing Data Warehouses Are Somewhere in the Process Between Exploring and Deploying Unstructured Data Analysis Solutions⁷

IDC predicts, “Over time it will become easier to deal with unstructured data as (1) more and more metadata is added to unstructured data, (2) structure is added to unstructured data, and (3) access systems provide structured views of both structured and unstructured data.”⁸

⁴Gantz, J., et al. (March 2007). IDC. The expanding digital universe.

⁵Feldman, S. (March 2007). IDC. Worldwide search and discovery software 2007-2011 forecast.

⁶Russom, P. (2007). TDWI. BI search and text analytics.

⁷Ibid.

⁸Gantz, J., et al. (March 2007). IDC. The expanding digital universe.



CHALLENGES DELIVERING MORE INSIGHT TO BUSINESS UNITS

As the group that implements and supports BI and data warehouse solutions, IT faces growing demands for analysis from business units.

Existing business systems cannot analyze textual content such as customer comments.

Departments Demand More Insight from Data

Business units look to IT for deeper insight into their rich data repositories. For example, suppose an emailed customer complaint escalates to a vice president. That executive wants to know whether the complaint is an isolated occurrence or a sign of bigger problems for the company. Turning to IT, the vice president asks if customer complaint insight can be analyzed using existing BI or data warehouse solutions.

Because there is no adequate way to examine unstructured data, IT reports that analyzing customer complaints is beyond the capabilities of existing systems. Current business systems can only correlate structured content. However, a great deal of unstructured data holds many of the answers business units are seeking.

Because data hidden in text is out of reach of existing business system investments, IT might examine departmental point solutions that can perform unstructured data analysis. However, implementing a different analysis solution for each division of the company creates great costs and support burdens for IT.


Unstructured data often contains insight that could strengthen or contradict business decisions.

Ideally, your unstructured data analysis solution should interface with your existing business systems investments, and enable all departments to leverage unstructured knowledge.

Business Units Make Only Half-Informed Decisions

Existing BI and data warehouse investments only take structured content into account. Unstructured data often contains insight that can strengthen or even contradict business decisions.

People express thoughts, feelings, and emotions differently in text-based communications. For example, email, survey comments, blog posts, and wikis all contain rich content. When this data sits outside the BI universe, a significant portion of insight is inaccessible for analysis.



Many different departments make critical decisions that could be improved with unstructured data analysis, such as:

- Marketing can analyze buzz that results from blog comments
- Operations can correlate issues in service records
- Compliance can monitor internal communications and identify situations that may violate standards
- Sales can understand events as they occur with major accounts

A brief examination of the history of text analysis provides further insight.

Text analysis began during the Cold War and took off following the 9/11 attacks in 2001.

BRIEF HISTORY OF TEXT ANALYSIS


In the 1950s, the Cold War forced the U.S. military to seek faster methods to translate German and Russian intercepts. Military intelligence groups turned to automated language translation solutions that quickly analyzed content and presented an English-language equivalent.

Fast-forward to the 1990s. Search engines such as AltaVista and Excite emerged, allowing English language queries on large amounts of content. The search capabilities of the Internet were not yet available inside the enterprise, despite the exploding amount of internal business content. If systems were available, they required specially trained personnel to find content.

By the late 1990s, wide consumer adoption of search filtered into the business world. Intranet solutions and appliances allowed basic enterprise search capabilities.

During this time, the U.S. government was applying text analysis to intelligence operations, seeking to understand the essence of communications, regardless of the language.

The 9/11 attacks in 2001 were a major turning point. Intelligence departments began sharing insight and needed robust tools to extract knowledge from large amounts of unstructured data. The U.S. government began leveraging commercial text analysis applications to identify trends and correlations.



Text analysis examines the linguistic structure of written content and extracts critical data that can be understood by computer systems.

By 2002, the financial industry faced heavy regulation that forced businesses to retain communications. Text analysis was used to help examine this large amount of data. Businesses began applying text analysis to identify problems early and mitigate risk. Text analysis became an early warning system of potential problems.

More recently the explosion of user-generated content in blogs and wikis and the massive growth of email are resulting in rich sets of unstructured data. Textual content is now important for sales, marketing, and many other corporate divisions.

GAINING NEW INSIGHT WITH UNSTRUCTURED DATA ANALYSIS

Designed to bring added insight to BI and data warehouse systems, unstructured data analysis products intelligently evaluate textual content and extract structured meaning.

BI consultant Seth Grimes explains, “Text analytics is about making human communications comprehensible to computers. Despite the inapt ‘unstructured’ label, textual documents have linguistic structure that is easily comprehended by people. Text mining tools apply linguistic (natural language) techniques to parse this structure and model it in ways that computers can understand.”⁹

Able to automatically identify critical elements within textual content, unstructured data analysis can quickly extract insight from text communications. For example, people, places, companies, dates, units of measure, and concepts can be extracted (see Figure 2).

Business Events: Repron Electronics, Inc. said Thursday it had acquired DigEquip Corporation's AssetWorks asset and desktop management solution. Repron appointed Michael Branca Chief Financial Officer effective July 16.			
RULE:	DETECTED:		
MERGER&ACQUIS.: BUYER::TARGET	BUYER:	TARGET:	
	Repron Electronics, Inc.	DigEquip Corporation's Asset Works	
EXECUTIVE JOB APPOINTMENT: COMP::PERS::POSIT	COMPANY:	PERSON:	POSITION:
	Repron	Michael Branca	Chief Financial Officer


Entity Attributes and Associations: The suspect John Smith is 6'1" tall. Mr. Smith works as a carpenter for WestCo. Smith went to Texas on vacation with his family and he met with Bill Jones at the park.			
RULE:	Detected		
HEIGHT: PERSON::MEASURE	PERSON:	MEASURE:	
	John Smith	6'1"	
OCCUPATION: PERSON::OCCUPATION	PERSON:	OCCUPATION:	
	John Smith	Carpenter	
TRAVEL EVENT: PERSON::PLACE	PERSON:	PLACE	
	John Smith	Texas	
ASSOCIATION: PERSON::PERSON	PERSON:	PERSON:	
	John Smith	Bill Jones	

Figure 2: Examples of How Knowledge Is Extracted from Text

“Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before, which helped them to gain more precision in fact-based decision making. After all, without representation from unstructured data, a data warehouse is a single truth, but not the whole truth,” notes a TDWI research report.¹⁰

⁹Russom, P. (2007). TDWI. BI search and text analytics.

¹⁰Russom, P. (2007). TDWI. BI search and text analytics.



Unstructured data analysis can extract a significant breadth of content from written communication out-of-the-box. Examples include information about:

- People, such as names, positions, and social security numbers
- Things, such as products, businesses, and financial indexes
- Time, such as dates, days, holiday, and time periods
- Quantities, such as money, sales, and measures
- Concepts, such as global piracy and policy issues
- Relations and events, such as the organization an individual works for

Additional list-based or pattern-based items can be customized for each business application. For example, a list of competitors or persons of interest can be added to the solution for analysis.

You can apply unstructured data analysis to many different sources of content, including:

- **Internal data**, such as email, customer surveys, Microsoft Word documents, Adobe PDF files, internal wikis, presentations, online forms, online chat, and notes fields in customer relationship management (CRM) or sales automation systems.
- **External data**, such as blogs, forums, wikis, analyst reports, news feeds, journals, patent filings, press releases, and competitor Web sites.


Designed to tap into existing BI and data warehouse systems, unstructured data analysis solutions bring insight from unstructured content into a structured environment. This new insight can be analyzed on its own or in concert with preexisting structured data.

Now business intelligence and data warehouse investments can analyze the full set of knowledge within the enterprise.

THE BENEFITS OF UNSTRUCTURED DATA ANALYSIS

Unstructured data analysis provides a wide variety of advantages to enterprises, including:

- Unlocks hidden knowledge that can be directly leveraged by existing BI systems
- Ensures important content is part of the business analysis and decision-making process
- Allows more precise fact-based decision-making

- 
- Quantifies facts previously unquantifiable, such as claims processing or blog comments
 - Transforms data warehouses from a single truth to the whole truth
 - Enables businesses to rapidly and consistently analyze large amounts of textual content
 - Allows IT to proactively offer business units new ways to analyze content
 - Reduces costs by automating the process of analyzing text, eliminating manual procedures
 - Results in better reports and trend analysis as business units can now examine the full spectrum of available content
 - Provides context so business leaders can identify events as isolated or part of a bigger problem
 - Enables businesses to proactively monitor internal and external data sources for opportunities or violations of standards


HOW UNSTRUCTURED DATA ANALYSIS WORKS

Consider the insurance industry. Claims for an auto collision include lots of written descriptions. Unstructured data analysis can examine certain types of accidents or a policyholder and associated location details such as the street name, house number, and town. When analyzed aggregately, a data warehouse solution could identify dangerous intersections or geographical areas prone to burglary.

Unstructured data analysis solutions can quickly target specific text sources for analysis.

In this example, the unstructured data analysis solution is tasked to examine a set of documents or another textual data source. To handle a large volume of documents, a taxonomy can be leveraged to identify the topic of each examined document. The system can also be configured to look for specific types of information, such as people, street addresses, towns, type of loss, and so on.

A sophisticated linguistic process identifies the language of documents as well as words, phrases, sentences, and paragraphs. Using grammatical analysis and semantic dictionaries, unstructured data analysis looks for patterns of words to identify items, such as people or place names.



Critical written content within a document is also summarized. For example, a multipage accident report can be summarized to a paragraph, which includes key aspects of the larger textual document.

The resulting content is structured data about the original unstructured data. All content analysis, such as topics, entities or facts, and summaries, is stored in a relational database, along with the location of the original document for reference. This new structured content can be leveraged by BI and data warehouse solutions.

WHAT TO LOOK FOR IN AN UNSTRUCTURED DATA ANALYSIS SOLUTION

When seeking an unstructured data analysis solution, consider the following important requirements:

- **Seamless integration with BI systems:** The ideal solution is immediately recognized by the BI system, eliminating the need for custom programming.
- **Highly customizable:** Look for a solution that can be easily customized to your business-specific requirements with an easy-to-use graphical user interface, eliminating the need for costly onsite analysts.
- **Interactive visualization:** Seek a wide range of visualization tools to help display unstructured data, including pie charts, histograms, heat maps and interactive dashboards.
- **Intelligent entity extraction:** Based on clues from the context of textual data, the ideal solution should be able to discover people, companies, dates and other aspects without reliance on lists or string matches.
- **Highly scalable:** As the volume of unstructured data increases, the solution should scale. Look for a solution that leverages distributed servers for large-scale unstructured data analysis.
- **Wide language coverage:** For multinational organizations, seek a solution that addresses all the languages of the business. Look for the ability to accommodate complex languages such as Chinese, Japanese, and Arabic.
- **Proven provider:** Only work with a company that has a proven track record working with Fortune 1000 businesses and leading government agencies. The company should be committed to long-term research and development and have a legacy of unstructured data analysis.

Seek a solution that effortlessly integrates with existing business systems.

THE BUSINESS OBJECTS SOLUTION

Designed to meet all the objectives outlined in this white paper, BusinessObjects™ Text Analysis provides organizations a complete view of the insight hidden in their enterprise data. Deployed at more than 400 global locations, organizations such as SAP, Yahoo, Microsoft, ConocoPhillips, Dow Jones, Oracle, and government agencies rely on BusinessObjects Text Analysis.

Taking its roots from the Xerox PARC research center, more than 20 years of development have gone into the technology. In addition, more than 75 patents in language processing and visualization stand behind BusinessObjects Text Analysis.

BusinessObjects Text Analysis is deployed at more than 400 global locations.

Built to integrate seamlessly with business intelligence and data warehouse solutions, BusinessObjects Text Analysis allows IT to better service the analysis requirements of business units.

Key capabilities include:

- Automatic extraction of key entities, such as business names, people and places
- Intelligent summaries of textual documents
- Categorization of documents into taxonomies based on topical discovery
- Wide-ranging support of different document types, including Microsoft Word, Adobe PDF, HTML, email, Microsoft Excel and RSS feeds
- Broad language support, including Spanish, Chinese, Dutch, Japanese, Arabic and dozens of other languages
- Integrates with BusinessObjects Enterprise and BusinessObjects Data Integrator applications
- Leverages service-oriented architecture for far-reaching system integration

Empower business units to make better decisions with BusinessObjects Text Analysis.

ABOUT BUSINESS OBJECTS

As an independent business unit within SAP, Business Objects transforms the way the world works by connecting people, information, and businesses. Together with one of the industry's strongest and most diverse partner networks, the company delivers business performance optimization to customers worldwide across all major industries, including financial services, retail, consumer-packaged goods, healthcare, and public sector. With open, heterogeneous applications in the areas of governance, risk, and compliance; enterprise performance management; and business intelligence; and through global consulting and education services, Business Objects enables organizations of all sizes around the globe to close the loop between business strategy and execution.

To learn more about BusinessObjects Text Analysis, visit www.businessobjects.com/unstructured/.

businessobjects.com