InfoTrends
A Questex Company
Focused | Responsive | Credible | Visionary

September 27, 2010 **Analysis**

Unlocking the Power of Content

Authors

Randy Dazo

Bryan Yeager

Chris Bondy

Published by

Dynamic Content Software Strategies

Production Workflow and Customized Communications Consulting Service

Abstract

Today's organizations are able to innovate faster, predict better outcomes, and make smarter business decisions by unlocking the insight within business critical information that comes in the form of unstructured or semi-structured content. With information constantly increasing, identifying trusted high-value content from incomplete, irrelevant, and outdated information is also greatly important. Many organizations are unaware that powerful content analytics solutions are available to tap into their content sources. This document outlines some key approaches to unleash the power of content, and describes the tools and solutions used to provide actionable insights from content.

For More Information

If you would like to order extra copies of this report, receive permission to use any part of the report, or be informed of upcoming market updates, reports, and related projects, please e-mail us at info@infotrends.com.

Headquarters:
97 Libbey Industrial Parkway
Suite 300
Weymouth, MA 02189
United States
+1 781 616 2100
info@infotrends.com

Europe:
3rd Floor, Sceptre House
7-9 Castle Street
Luton, Bedfordshire
United Kingdom, LU1 3AJ
+44 1582 400120
euro.info@infotrends.com

Asia:
Hiroo Office Building
1-3-18 Hiroo, Shibuya-ku
Tokyo 150-0012
Japan
+81 3 5475 2663
info@infotrends.co.jp

Table of Contents

Introduction

Today's organizations are increasingly being tasked with improving revenue generation at a time when markets are challenged and cost reductions have become necessary to meet budgets. While technology can certainly improve efficiencies, true innovation will come from companies utilizing the information they already have to create new business opportunities, reduce costs and risks, improve business decisions, or streamline processes. These approaches can ultimately increase sales and customer satisfaction while reducing bottom-line costs.

Over the years, organizations have collected an abundance of information that is mostly unstructured. This information consists of documents, e-mails, wikis, chat logs, blogs, and Web forms. In addition, structured information continues to increase, but only reveals, in many cases, a portion of an entire story. True innovation and insight has been demonstrated when organizations leverage structured and unstructured information that can be trusted, exploited, and then analyzed.

For example, in the healthcare industry, unstructured patient notes and structured clinical data can be co-analyzed to improve the accuracy of diagnoses and contraindications. Unstructured patient notes have a wealth of actual accumulated patient information that rarely gets synthesized and analyzed with structured data for future use.

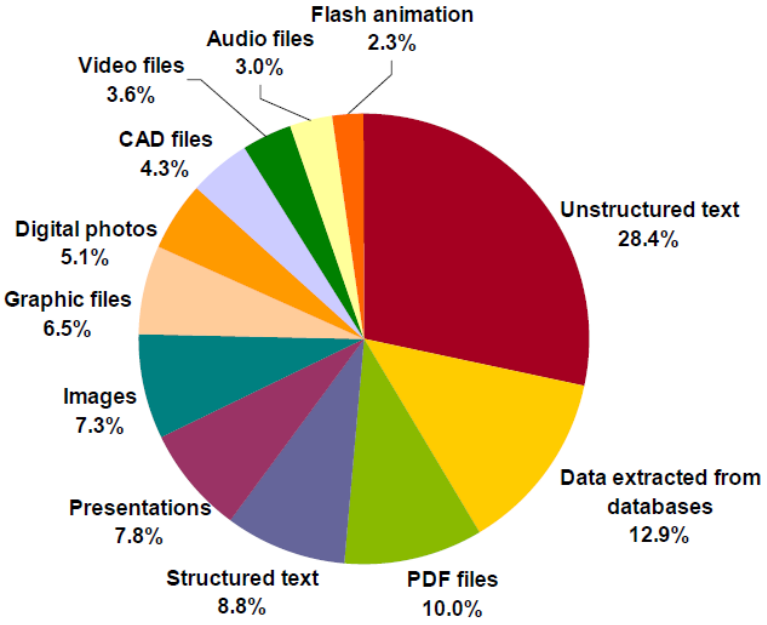
Organizations need to know that an analytics approach can be carried out to successfully leverage the wealth of information that exists in most organizations today. Content analytics should be viewed as a new approach to existing infrastructures that can help bring out the value of this information. This can be addressed by looking at the current areas where information exists, structured and unstructured, internal and external to a company that can help drive better business decisions. This can happen within an individual group or throughout an organization as long as decision makers recognize that there is hidden value within their information.

From Chaos to Innovation

Information within an organization has exploded and expanded its footprint to a variety of document types, ranging from e-mail to PDFs and digital media. The challenge with content today is that it resides anywhere within the organization: in a variety of repositories, disparate file systems, and even on individual hard drives. In addition, with content doubling year-over-year, storage requirements are growing and the burdens of managing multiple repositories are increasing. This makes the unification and delivery of trusted content to the business very difficult.

As shown in Figure 1, the type of business content that resides in organizations is fragmented, ranging from multimedia files and slide presentations to PDF files and unstructured text. This type of variety presents an interesting challenge to those trying to gain knowledge from their content repositories as they strive to better manage this content.

Figure 1: Business Content by Type (N=202)[1]



The following are three key approaches that can help an organization take hold of their information and begin to take advantage the power within:

- **Know your content**
  - Dynamically analyze what you have, decommission the unnecessary, and preserve the content that matters for compliance or eDiscovery needs
- **Trust your content**
  - Manage and govern content in trusted repositories, not in suspect environments, enabling confidence in your content
  - Create and manage 360 degree-trusted content views to enrich master data by connecting to enterprise content
- **Analyze and leverage your content**
  - Interactively discover content to derive unexpected business insights and take action with content analytics
  - Exploit content analytic insights by enriching existing enterprise content management repositories, improving business intelligence and predictive analytics, as well as tailoring for industry and customer specific scenarios

[1] Franke, Maziarka, and Duek. *Multi-Channel Communications: The Content Publishing Workflow Challenge.* InfoTrends, Inc.

Gain Control with a Content Assessment

A comprehensive assessment of all repositories, content, and the business strategy is a key component of discovering usable and trusted content. More than ever, organizations are aware of the soft costs tied to information access and employee productivity. Siloed content repositories, numerous user interfaces, and an inability to granularly search across the enterprise have stifled efficiency. Users are often forced to jump between multiple application windows as they search, cut, and paste information into a form that is relevant and useful. The information explosion has exacerbated this problem, and knowledge workers today are lucky to find important content in a timely manner at all.

The exponential growth of electronic information in many organizations today has become one of the biggest challenges to managing and controlling content. Because 80% of information in most enterprises is unstructured content that is trapped in siloed repositories, it constrains their ability to manage and efficiently leverage relevant and necessary information that can be used strategically to the company's advantage. Unfortunately, quite a bit of this information is also deemed unnecessary and irrelevant, most companies take the "keep everything" approach because of the daunting task of having to sift through the enormous amounts of information in a variety of repositories. This can pose costly compliance risks and negatively impact IT budgets for having to maintain infrastructure and hardware, as well as forcing companies to incur unnecessary associated administrative costs.

Companies require tools to confidently help them assess their content to build criteria and a decision plan to properly manage their information. Assessing information using content analytics can help aggregate, correlate, visualize, and explore unstructured content. It is the first critical step to alleviate the pain of the information explosion by evaluating whether it is necessary or unnecessary to the business, and enable informed decisions about business value, relevance, and disposition.

Analyst firms studying the productivity of knowledge workers have quantified the costs of searching content. Although the methodology of these studies varies, results indicate that the average knowledge worker spends 6-10 hours per week (approximately 1-2 hours per day) searching for information within the enterprise. Based on a salary of $60,000 for a knowledge worker, this translates into organizational costs of $9,000 to $15,000 per year per employee. If an organization with 1,000 knowledge workers can save 50% of the time spent on search-related matters, this could generate up to $7.5 million in worker-related productivity per annum. Of course, these figures become even more significant when considering higher-paid workers and those employees that conduct an above-average number of searches.[2]

Whether the need is to meet compliance and eDiscovery deadlines, analyze historical data for present business decisions, or empower analytics-driven business applications with trusted content, information access and analysis will define the next generation of competitive organizations. These data-driven businesses will acknowledge the information explosion and deploy an enterprise-wide set of tools to identify valuable content while decommissioning insignificant content.

Organize and Manage Trusted Content

Organizations need to automatically organize and tag content to improve the management and compliance, as well as gain leverage for additional business insight. Solutions, such as rules-based classification tools,

[2] IDC, Delphi Group, et al. Based on a 40-hour work week.

can be used for general organization that is based on pre-defined conditions. There are new advanced classification solutions, however, that can help organize unstructured content and automate daily, content-centric decisions by analyzing the full text and context of documents to correctly sort out useful information and then categorize and tag it accordingly.

With many unstructured documents, just looking at certain keywords and rules is not enough to determine if a document is business-centric or can be trusted. For instance if the word "sue" appeared in an e-mail, the word "sue" could be surrounded by context about a lawsuit that would need to be tagged for the legal department or it could be just be a reference to someone's name in a personal e-mail. In this case, typical rules-based classification solutions may be setup to tag any form of content with the word "sue" for legal, which would include critical and non-critical content and would probably have to be manually sorted out later. Using advanced forms of classification can solve this problem because it can contextually understand the difference between the usage of "sue" vs. "Sue" and classify it accordingly. Being able to leverage context-based and rules-based classification can establish the integrity required for trusting information.

Another key part of organizing information is to create and manage a 360 degree, single view of trusted entities by connecting valuable meta data from enterprise content to master data management repositories; essentially linking trusted metadata between content and data . Critical information to an organization, such as a customer's name, can often times be referenced or entered in various forms (e.g., with nicknames, different capitalizations), essentially having the same customer with multiple entities. Being able to tag these different records as a trusted single entity can help gain control of data, reduce information errors, and eliminate duplicate data—all of which will let organizations meet growth, revenue-generation, and cost-reduction goals

UIMA at a Glance

Unstructured Information Management Architecture (UIMA) is an open source project of the Apache Foundation that was originally pioneered by IBM and approved as an OASIS standard in March 2009. UIMA is a pipeline processer with frameworks or APIs that can perform text analytics functions, which is a key component to performing advanced content analytics on unstructured data. These functions include:

- Tokenizing words into structured sentences and paragraphs for further analysis
- Deriving stems from words in multiple languages
- Pinpointing and classifying common expressions, such as e-mail addresses and phone numbers
- Accessing dictionaries to annotate, decipher, and derive meaning from unstructured information
- Customizing annotators to obtain specific desired results

UIMA can be utilized by existing text processing tools, including the General Architecture for Text Engineering (GATE); OpenNLP; and numerous commercial analytics tools, including those from IBM.

Handling Unstructured Information

Unlike the aforementioned structured text and data that typically resides in a database, unstructured information presents a unique challenge when separating valuable content from information that can be archived or discarded. Instituting and following best practices with Enterprise Content Management goes a long way to help tag and categorize unstructured information. Nevertheless, sometimes things fall through the cracks even with a robust process in place—content is duplicated and classification can be incomplete, totally missing, or even applied wrong. While classifying and categorizing content as it enters

a system can be easily accomplished as part of the workflow, the process of finding and correcting errors, de-duplicating, and decommissioning non-valued assets can be demanding. Even so, this process is necessary if enterprise organizations want to cut down on rising storage costs.

Fortunately, great technological strides in processing and analyzing unstructured information are bringing automation to a burdensome, costly activity. This can be a huge benefit to enterprise organizations. Unstructured information varies from PDF files and slide presentations to e-mails and logged chat conversations. Metadata provides some insight into the details of unstructured information, but it can only go so far in its description of that content. More advanced techniques can be utilized to accurately parse and analyze the textual elements of unstructured content, which can then be used to determine the value of the information and the appropriate actions that need to be taken. A prominent example of technology that can enable content analytics is the Unstructured Information Management Architecture (UIMA).

UIMA is an architecture upon which sophisticated content analytics solutions can bring structure to unstructured information and enable the use of powerful analytics to gain deep insight and knowledge on content that could not be achieved through manual analysis. By giving this content structure and applying analytical technology to it, a vast number of new possibilities emerge. Content analytics can be used to synthesize structured and unstructured information from disparate repositories to provide a complete picture of a particular entity. For instance, a police department could utilize content analytics to bring together various types of information on a criminal suspect that comes into the station, such as outstanding warrants, fingerprints, mug shots, police reports, case files, and prison records. This application can enable officers to get a jumpstart on a case by quickly providing a complete assessment of the suspect.

In a similar scenario, a police department could employ content analytics when attempting to solve a case. Analytic technology is typically used to find patterns, correlations, anomalies, and outliers within a data set, often through the use of visualization. Content analytics exploits those same utilities and enables other people, like crime analysts, to bring in trusted unstructured content that may provide new clues.

By gaining this type of understanding, organizations can more easily synthesize information and gain previously-unseen insight to help make informed decisions.

Analyze and Leverage the Content

As seen in the previous examples, organizations can leverage their trusted and organized content efficiently and more effectively to predict outcomes, make better and faster decisions, or increase opportunities. These occasions can be broken down into the following solutions:

- Improved business process or case management
- Increased insight and predictive solutions
- Business intelligence solutions
- Legal and regulatory compliance analytics

Further examples include:

- In the insurance industry, it is important to be able to detect fraudulent claims before payment occurs.

- In product development, businesses must understand what customers want in new features or capabilities.
- Law enforcement personnel need to be able to predict crimes as well as apprehend criminals at the time of their offense.
- In financial services, unstructured communications and structured market data are used to automate and inform highly agile trading strategies.
- In the distribution and logistics industries, rapid access and understanding of information fuels just-in-time delivery and supply chains.
- In the energy sector, real-time information collected from the U.S. "smart grid" will drive automation of energy production and routing during peak periods, as well as preventing brownouts and grid meltdowns during crisis situations.
- In the travel industry, internal and external data sources are used to intelligently price offerings to maximize profits.
- In the advertising/marketing industry, data from Web, print, e-mail, and social media promotions are being used for multi-channel and highly-targeted campaign management as well as lead generation.

An Analytics Approach

Content analytics does not necessarily rely on a certain organizational infrastructure to be put in place nor does it require a strict strategy to implement. Organizations can take advantage by deploying several approaches simultaneously and do not require a strict path for engagement. Because of this, organizations should look at deploying analytics approaches based on their business needs. A small department or group can implement a content strategy simultaneously with the business, starting an enterprise-wide engagement. As with any new approach, there are several things to bear in mind to achieve the best results:

- A combination of structured and unstructured information: bridging the gaps will provide more insight and yield greater outcomes
- Gaining control of content: understand where content lies throughout the organization and begin to take control to be able to leverage it
- Classification and tagging of multiple similar entities that can be viewed as a single trusted entity
- Organization and auto-classification of information can enable trusted content to be readily usable and accessible
- A process through technology that supports the governance and compliance initiatives as well serve the current and trusted content, preventing human error
- A modular approach can manage immediate needs (such as e-mail, file shares, and SharePoint archiving), while also being strategically flexible and scalable to include additional capabilities or content sources

InfoTrends' Perspective

The exponential rise in content stored by organizations drives up costs such as storage and litigation, and yet is rarely leveraged to deliver new business insight. While traditional Enterprise Content Management solves some of the problems with storing, classifying, searching, and categorizing unstructured data, the process of determining what to save, archive, or decommission remains a laborious manual process. Pushed by the need to innovate, businesses need to cost effectively control content growth and identify new business opportunities by gaining new insights into their information.

Unstructured and structured information can be harnessed, aggregated, and analyzed to provide deeper insight in a wide array of areas. By deploying content analytics approaches, organizations can increase their agility in a number of areas. They can develop a holistic view of a customer, quickly catch product defects, detect fraud, and ultimately make smarter decisions that positively impact the bottom line. Even outside the realm of business, content analytics can be used to write better legislation or even fight crime more effectively.

Ultimately, instituting content analytics is the next logical step for enterprise organizations. Whether companies are trying to rein in mounting storage costs or make smarter, predictable, and more informed business decisions, content analytics technology opens the door to a world of new opportunities and insight by unlocking the power of content.