IDC
Analyze the Future

WHITE PAPER

Content Analytics and the High-Performing Enterprise

Sponsored by: IBM

Susan Feldman
Hadley Reynolds
David Schubmehl

December 2012

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

IDC OPINION

Information has become the lifeblood of today's organizations. We need to know what our customers are saying or our competitors are planning. Above all, we need to mine all of our information to get a full view of our business. This vital information is contained in text — the documents, the email messages, the comments in the media or on social forums. To uncover the known and the unexpected, we need to be able to search for and analyze text and, by extension, other content types, including voice, sound, and images. Search engines and content analytics are a collection of technologies that have been developed to understand, retrieve, and mine content. They add the "why" of an event or a trend to the "who," "what," "when," or "where" that structured data provides.

The Business Imperative

Global business today faces unprecedented challenges in managing the quantities of information that flow in and around all organizations every minute of every day. IBM's recently published survey of over 1,700 global CEOs (*Leading Through Connections*, May 2012, IBM Institute for Business Value) shows that CEOs expect to gain a competitive advantage by using new technologies that mine and analyze today's massive information flows.

In IBM's research, CEOs in companies that "outperformed" their peers were distinguished by their ability to access data, draw insight from that data, and *translate that insight into action*. One CEO stated: "*Survival skill 101 for the next five years will be deriving insight ahead of peers.*"

Traditional software tools that the enterprise uses to derive insight from digital information include database technologies, enterprise applications, search, and business intelligence (BI). These powerful tools mark the boundaries of IT practice today, yet they have not solved the challenge of turning increasingly large and rapid data flows into insight for most employees and executives in most organizations. Innovative technologies such as content analytics are already being adopted by outperforming organizations to solve this problem. To the solid reliability of database applications, they add agility, speed, and insight into the 85% of information that is not structured so that business professionals can gain insights, answer questions, and make decisions.

What Is Content Analytics?

Content analytics applications process content to extract information building blocks — ideas, names, time, events, and even opinions. They forage for relationships among these building blocks — such as cause and effect, acquisitions and mergers, or side effects of drugs. Content types include photo captions, blogs, news sites, customer conversations (both audio and text), social network discussions, faces, maps, multimedia resources, and conventional text documents. The technologies used for this purpose include entity extraction, relationship detection, sentiment analysis, reputation management, trend analysis, affinity, recommendations, face recognition, speech analytics, visualization, and industry-focused custom analytics. Depending on how they are combined, they are the foundation for such applications as:

- "Voice of the customer" to analyze call center data and other customer communications
- Crime and criminal activity analysis and detection to support law enforcement activities
- Competitive intelligence in a broad range of settings
- Product launch analytics to gauge consumer reactions to a new product
- Fraud detection in financial services, government benefits administration, insurance, and many other settings,
- Many other applications specific to domains from healthcare to horticulture

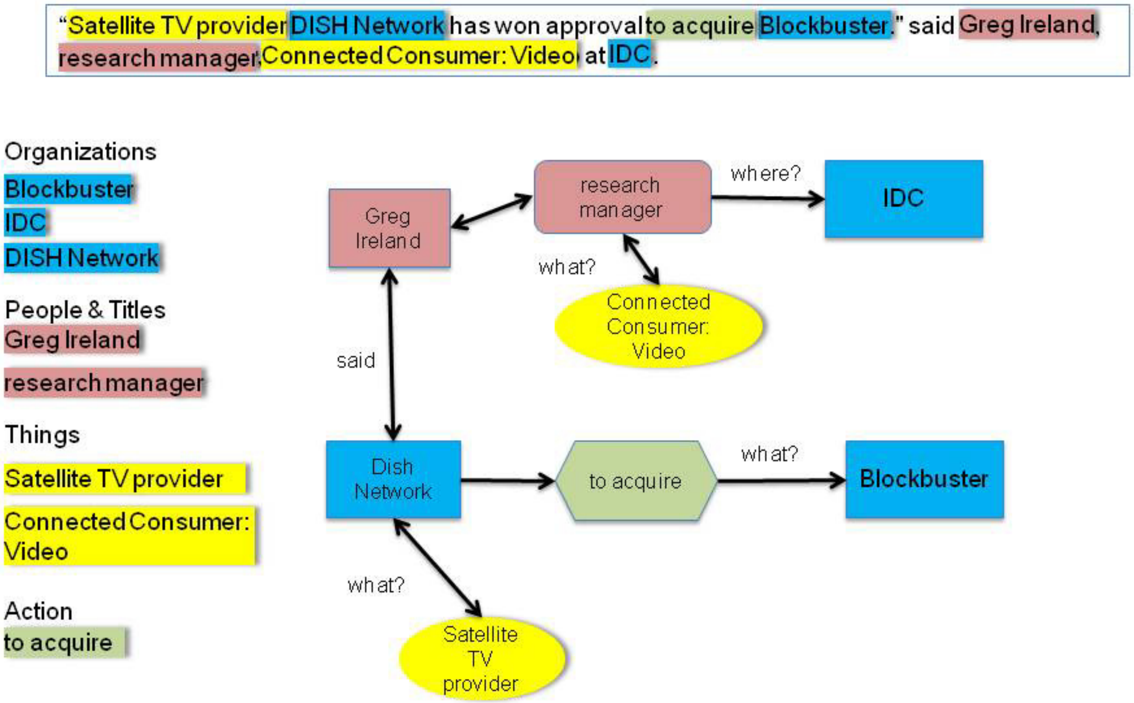
Content analytics enhances existing software applications and is the foundation for a new set of software applications that are aimed at understanding and mining text for new types of business intelligence. Like the data used in traditional business intelligence applications, elements extracted by content analytics can be sorted, compared, or plotted over time to spot trends or report on the status of recurring business processes or topics of interest. Unlike BI, content analytics monitors the predictable and the unpredictable actions and interactions of people to alert businesses to the unexpected. Content analytics can be used to spot emerging trends, new markets, or trouble spots in customer relations. At its most advanced level, content analytics can answer questions directly. Combined with existing applications, this new technology foundation adds a dimension of language understanding to enhance and extend business intelligence, search, or call center applications to improve productivity and present a clearer view of the business. For these reasons, this segment of the search and discovery market is showing a high rate of adoption and use.

Content analytics tools build on approaches developed in enterprise search and business intelligence technology for more accurate information access, greater ability to navigate information collections, and integrated views of structured and unstructured data. These capabilities may serve as an overlay technology suite to improve the performance of search and BI applications. They are a key element in unified information access applications, which respond to the business' need for highly specific and efficient decision support environments or for customized research engines.

A simple example of content analytics can be seen in the analyzed sentence in Figure 1. In this particular case, we know that "Blockbuster" is a reference to a company, not a popular movie. We also know that "Ireland" is not a place, but the name of a research manager at another organization called "IDC." Content analytics provides the tools to identify people, places, things, and actions so that links and correlations can be made for research or discovery by tools such as enterprise search systems or business intelligence tools.

FIGURE 1

Contents Analytics Example



“Satellite TV provider DISH Network has won approval to acquire Blockbuster." said Greg Ireland, research manager, Connected Consumer: Video at IDC.

Source: IDC, 2012

The Role of Content Analytics

Today's big data environment adds to the complexity of information analysis, but big data also creates opportunities to understand customers, patients, citizens, trends, and ideas as never before. Precisely because of the volume of big data, we can increase the certainty of our analyses, and we can create microsegments of populations that will enable us to individualize ecommerce or to diagnose patients.

Most of the data that swirls around the organization is unstructured — text, video, images, social media, Web pages, mobile, and audio communications. Unfortunately, most of the tools IT has available to analyze these large quantities of data are designed for the structured data world, where information is carefully organized in rows and columns and structured in predefined schemas. Unlike databases, text and rich media — particularly rapidly changing streams of such information — cannot be captured, analyzed, and managed effectively using structured data tools such as databases and data warehouses.

Content analytics straddles the line between the unstructured and the structured information worlds. It extracts the elements of meaning from unstructured information and presents them in a more structured format so that they can be combined and analyzed in concert with structured data. Content analytics is also used to extract meaningful information from databases and can normalize across both databases and collections of content in order to find relationships that cross the boundaries of the source collections. These technologies categorize and tag documents to emphasize what they are about. They extract names of people, places, products, and things — entities — as well as time, opinions, sentiment, and geographic location — and add this additional information as metadata to the search index. The results are striking: Search is improved because the additional metadata helps the search engine return more relevant results. The tags — entities, time, location, and topic — are used as facets so that users can explore a collection by browsing rather than searching. The structured output from content analytics can now be added to traditional BI tools to be used in more comprehensive business analyses. The results fuel trending software to find trends in stocks, disease patterns, or sales. They make it possible to add automatic question answering to online self-help sites, and they alert executives to surprises — both pleasant and unpleasant — that are found in unstructured information. Customer complaints can be mined to identify the top issues and to head off problems with products. Content analytics is used in eDiscovery and in product development. In other words, it's the missing piece in analytics.

The content analytics operation has been likened to an information refinery, in which intelligent processors, referred to in the industry as annotators, extractors, or recognizers, add new levels of meaning or perspective to the index in addition to the original text. These new data elements create a new set of lenses for a virtually unlimited number of decision areas, which business implementers can tune to their particular areas of interest. Business decisions increasingly require executives to review data from both structured and unstructured sources; content analytics provides a bridge between the two. It brings unstructured information for the first time into the analytics practice, and it enables managers to consider integrated views of all relevant information for analysis, discovery, and decision making.

Most organizations have taken the first step toward taming their text by deploying search engines. But search can only go so far: It answers questions, but it can't explore without a query. Yet, information exploration is what yields the surprises that are often the most valuable result of mining information. IDC believes that although installing enterprise search can deliver immediate benefits in understanding collections and flows of unstructured data, organizations that want to maximize the return on their investment in data and text collections need to add content analytics to extract the most value from both unstructured and structured information resources. Intelligent content analytics allows an information system to:

- Categorize, analyze, tag, and extract information building blocks from text and other unstructured as well as structured forms
- Enable information discovery and exploration through entities, events, and relationships
- Present more accurate, contextual search results
- Deliver BI-like visualization of trends in the underlying content and data sources

Content analytics, while certainly robust and well-established, is relatively new to most enterprises. Investment in content analytics is growing rapidly as various content analytics–based applications become more widely adopted. In addition to improving search by adding concepts, entities, and other metadata, content analytics is widely used today in "voice of the customer" applications, in eDiscovery, in pharmaceutical research, and in government intelligence. The financial industry is adopting content analytics for fraud detection as well as to recommend investments to its clients.

Content analytics software offers a significant value uptick (e.g., over enterprise search technology) directly out of the box (e.g., by creating browsing interfaces to accompany search results). For more complex uses, with higher value, an enterprise-class content analytics practice calls on a toolbox of multiple complex technologies that organizations will want to tune to their own company business environment and the information domains relevant to their products or services. Organizations that seek a more precise understanding of their business will want to add terminology that is specific to their business goals and decision environments — taxonomies, categories, rules that initiate processes or actions. Accessible and efficient content analytics modeling environments are a requirement for outperforming organizations today. These tools will create the new class of innovative analytics applications that content analytics is capable of driving.

CONTENT ANALYTICS IN PRACTICE: PROMISE AND PITFALLS OF CONTENT ANALYTICS

Today, most organizations must accommodate change, and that means discovering emerging trends that might affect the business, understanding the nature of the impact, developing a strategy for change, and implementing a change management plan — quickly. Content analytics offers enterprises a number of advances over traditional tools, particularly in the areas of discovery and understanding. At the same time, there are inherent challenges in implementing these new systems. In the sections that follow, we examine a number of dynamics that can influence the quality of the outcomes from a content analytics application and highlight practices we have observed among successful practitioners.

Text Is Not Data

Text is not data in the type of information it contains or in its format. Data can contain the "what, where, and when" of an event, but text often contains the "why" and the "how." Text is ambiguous and unpredictable, able to state the same idea in multiple ways. Because content analytics is designed for language instead of numbers, it is probabilistic rather than deterministic. It allows a certain elasticity in matching meanings to accommodate synonyms and alternate phrasing of an idea. This statistical flexibility enables content analytics to discover expressions in text that carry similar meanings, not merely the exact matches to a database item or search term.

Content analytics tools require a different kind of processing than that typically used in conventional database or data warehouse applications. In the structured data world, data is organized into predesigned schemas that are designed to answer predetermined questions. In the fuzzy world of text, information nuggets — concepts, names, events, relationships, sentiment, location, time, etc. — are extracted into a flexible index and readied for combination and presentation at the time a human user performs a search or asks a question. Content analytics applications routinely present opportunities for users to discover entities, relationships, or correlations contained in the text that might never have been found in prestructured data.

In the structured data world, business intelligence solutions are increasingly used to discover suggestive patterns and relationships in the data. In the unstructured world of text, content analytics is the linchpin for this kind of discovery and analysis, using probabilistic tools such as fuzzy matching and machine learning to identify patterns and establish relationship links without the need for a predefined schema.

More Data Means More Options

More data is often better. In order to determine if a change in data is a trend or an accident, organizations need more data to confirm a hypothesis. The difference between 5 people and 50,000 people purchasing pet rocks can help a company make manufacturing and distribution decisions. More data gives early adopters an information advantage, allowing them to detect trends or threats before their competitors do.

Unstructured data has posed a particular challenge, both because it represents at least 85% of all data present in the enterprise and across the Web and because it has largely been opaque to the structured data-oriented processes of business analytics. Most organizations have attempted to implement strategies to reduce the amount of data they deal with, viewing the constantly expanding universe of data as a threat, not an advantage.

Content analytics can be a part of the solution to turning the data tsunami into an enterprise advantage. It does this by rendering the data in text documents into database-friendly formats, which conventional business analytics solutions can read and leverage for more comprehensive insights for decision makers.

Preparing for Content Analytics

Use content analytics tools to cleanse your text data

Determine what level of cleansing is "good enough" to get started

Data is never pristine. It needs to be curated, cleaned, and normalized. In the world of unstructured content analytics, dirty data can yield incomplete answers in analytics and search processes because of:

- Misspellings and alternative spellings
- Ambiguous references
- Changes in terminology
- Errors in facts
- Obsolete data
- Inaccurate data

The first job for content analytics implementation teams is to learn the data quality troubleshooting tools that content analytics software provides. Cleansing data should not be a completely manual effort. Advanced content analytics solutions offer tools to address the data quality issue. They can ferret out misspellings and alternative spellings, ambiguous references, etc. These tools offer multiple statistical methods, as well as conventional human test and review procedures to help discover and address dirty or incomplete data in an efficient manner.

Second, be aware that your text data will never be 100% accurate. A "dirty little secret" of data quality is that even manually entered or cleansed data is often only 70–80% accurate. The accuracy of automatic categorization and tagging techniques usually starts at this level and can be improved to 90% or 95% with customization over time. While content analysis techniques are never 100% accurate, the discoveries and correlations that can be made with even 80% accuracy is a great improvement over no insight into the information — the common situation in most organizations today. Organizations need to determine the minimum level of accuracy necessary to provide useful results.

Customer information, in particular, is an area in which content analytics is making a major contribution to business insight. Today's social media environment is an example of a big data challenge to the enterprise, in that important commentary from users about a company and its products is being generated every day in the growing collections of social data around the Web. Using content analytics software with sentiment monitoring, customer relationship and marketing managers can now mine not just calls to the call center but also emails to support sites, chat sessions with customers, responses on customer survey forms, and social media posts on services such as Facebook and Twitter to develop a truly integrated view of customer sentiment. This is an example of big data offering big insight for firms that take advantage of developments in analytics.

Influence of the Domain on Content Analytics Quality

Just as you wouldn't want a grocery clerk speculating about your chest x-ray, it's important to educate a content analytics system with knowledge about the company, the industry, the products, and the people that are most salient to the business. The accuracy of any content analytics application improves if it is primed with the terms, concepts, and knowledge that are part of any industry or profession in any given application environment, the general field or domain, the particular type of business, The specific details of the organization setting play a role in establishing the kinds of problems, issues, terminology, and conceptual relationships that are most relevant within the content.

Domain context is very important to the accuracy and impact of a content analytics system. An application in the healthcare environment, for example, will require an entirely different strategy for metadata, markup, analytics, and presentation than a content analytics application in the aircraft maintenance business. Even within the healthcare domain, content analytics applications designed to serve frontline caregivers will vary dramatically from those designed for hospital management or the payer-side professionals in the insurance industry.

Industries, organizational structures, professional roles, and even specific task contexts can require highly differentiated, domain-specific term and entity definitions, structures, and relationships. Content analytics application developers can incorporate all these elements in a system design that supports users' knowledge of their profession and how it operates. In contrast, generic systems force workers to work around or reinterpret the results.

Content analytics software helps application developers embed domain expertise in the form of dictionaries, taxonomies, or other knowledge bases by extracting key elements of meaning from the content. Core approaches available include:

- Automatic clustering routines to expose conceptual, entity-oriented, or linguistic patterns in the data

- Automatic and/or rules-driven taxonomies for classifying items in the data in a way that reflects domain knowledge structures

- Term lists, vocabularies, and thesauri to distinguish terms and phrases with domain-specific significance

- Lists of products, regions, offices, dealers, partners, competitors, or other entities and relationships of special interest in the domain

- Specifications for types of relationships such as mergers and acquisitions; initial public offerings; joint ventures; diseases, pharmaceutical antidotes, and causes of side effects; inventors and patents; criminals and crimes; etc.

All of these capabilities improve an information system's ability to understand the meaning of the data in the same way that ongoing professional experience makes industry experts more knowledgeable and valuable.

CONTENT ANALYTICS COMPONENTS

Content Analytics Modules

Content analytics relies on a series of modules that extract meaning and structure from the multiple elements that make text understandable to a human. Each of these modules can be used separately to improve the findability and discoverability of information. Some of the most common modules are:

- Linguistic analyzers find stems of words and tag the text for parts of speech.
- Entity annotators recognize the names of people, places, or things.
- Concept annotators or categorizers determine the major topics of each sentence, paragraph, and document.
- Relationship annotators determine the type of action being described and the relationships that exist between the entities and concepts that are acting or being acted upon.
- Date annotators determine the time when an action takes place.
- Sentiment annotators determine whether the opinions expressed are favorable or unfavorable regarding a specific entity.
- Geographic annotators associate mapping coordinates with any geographically relevant entity or concept in the text.

The list of annotators available from products "out of the box" is constantly expanding. Most implementations, however, will need to create one or more custom annotators to adapt the application to their specific use and the vagaries of their content.

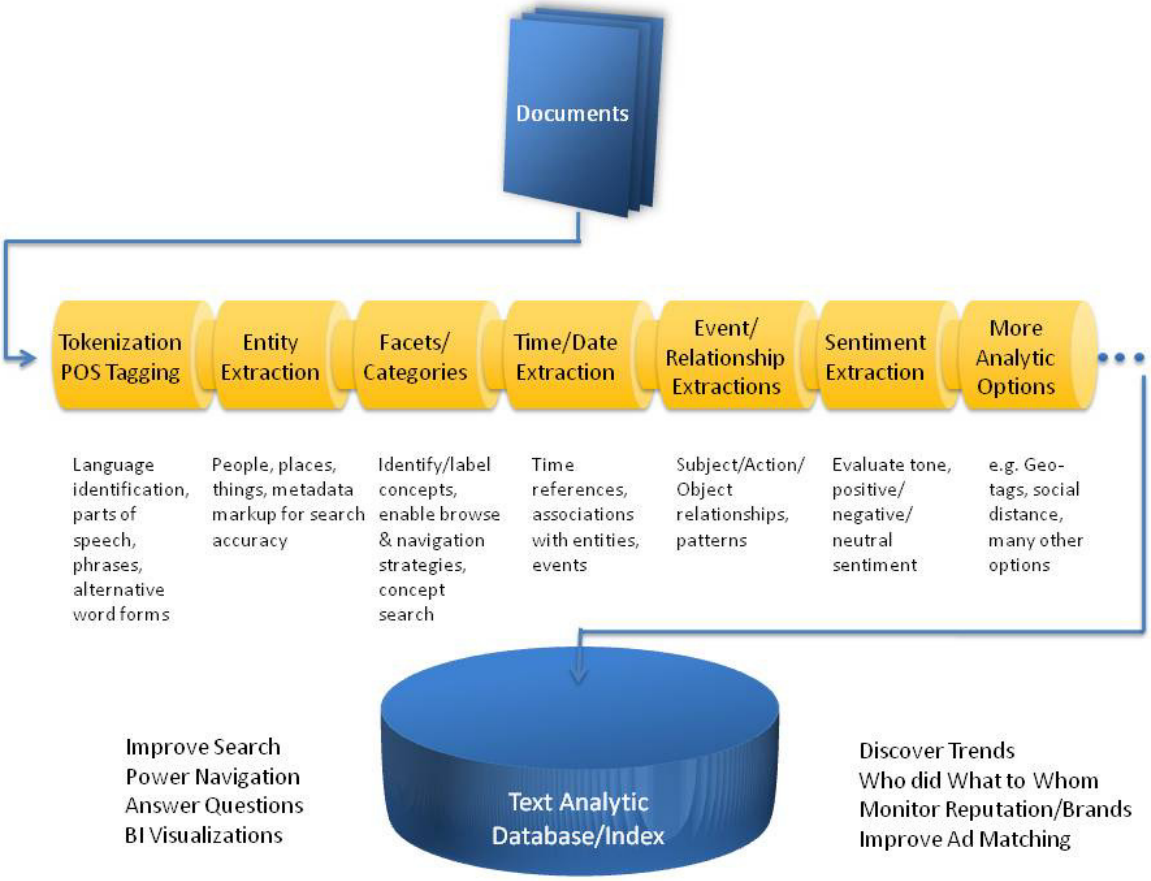
The Content Analytics Pipeline

In content analytics processing, a pipeline architecture has proven to be the most practical approach for software product suppliers as well as implementers, as can be seen in Figure 2. Because many kinds of components are typically mingled in a particular application, and because most of the components call on the same content analytics "primitives" in order to function, a pipeline approach steps the content through the tagging and extraction process in a logical and efficient progression. The first stages establish important linguistic data (e.g., language, parts of speech) that later stages (e.g., categorization) leverage, while later stages may branch to highly specialized routines that call on the data previously extracted (e.g., sentiment analysis).

Each content analytics component adds a distinctive value to the implementation as a whole, whether the end product will be delivered in search, content analytics, or business analytics applications. In the sections that follow, we identify major elements of the content analytics pipeline and point to the advantages of each. The order of the sections reflects the typical processing order in a working application.

FIGURE 2

Content Analytics Pipeline


Documents
Tokenization POS Tagging
Entity Extraction
Facets/ Categories
Time/Date Extraction
Event/ Relationship Extractions
Sentiment Extraction
More Analytic Options
Language identification, parts of speech, phrases, alternative word forms
People, places, things, metadata markup for search accuracy
Identify/label concepts, enable browse & navigation strategies, concept search
Time references, associations with entities, events
Subject/Action/ Object relationships, patterns
Evaluate tone, positive/ negative/ neutral sentiment
e.g. Geo-tags, social distance, many other options
Text Analytic Database/Index
Improve Search
Power Navigation
Answer Questions
BI Visualizations
Discover Trends
Who did What to Whom
Monitor Reputation/Brands
Improve Ad Matching

Source: IDC, 2012

Language Recognition, Tokenization, and Morphological Analysis

Identifying language is the first step in the content analytics pipeline. Without knowing which language a document is written in, it's difficult to understand its meaning. Any content analytics or search application performs operations such as language identification, tokenization (segmenting the letters into words), and identifying word stems before moving on to more advanced analysis. Increasingly, applications must handle a variety of languages, and this is a requirement that should be established before acquiring content analytics software to ensure that content across the spectrum of global languages can be processed.

Syntactic Analysis

Linguistic analysis is the core component of any content analytics system. This component analyzes language at the syntactic or grammatical level, looking for the role that the words play in a sentence. Some content analytics applications identify only basic forms, such as nouns and verbs. Others break down words into their roots and variant endings; recognize the parts of speech of every term within the text; and identify phrases and associate them in their correct position in a "sentence diagramming" procedure. The result of this processing is a metadata markup of all documents in a collection. Phrases can be used for faceted navigation. If the content analytics application has fully analyzed a sentence, then the relationships among the entities can be used to determine cause and effect, or definitions of terms. This deep syntactic analysis is a requirement for event extraction.

Practical Pointers

Content Analytics Processing Standards

Leverage content analytics software based on standards

Virtually all content analytics applications use proprietary APIs to integrate functions. Many organizations, therefore, have faced severe challenges attempting to construct integrated analyses from applications from different vendors. Standards efforts have been under way to address these incompatibilities. IBM took the lead in 2005, with support from the U.S. government, by developing UIMA, the Unstructured Information Management Architecture. By 2006, IBM released the code for UIMA to an open source process, which is currently managed by the Apache Software Foundation. Subsequent open source efforts offer their own "standards," such as Apache's OpenNLP or the University of Sheffield's GATE or elements of the Apache Lucene/Apache Solr distribution. Numerous organizations work with these systems as well as with UIMA. Content analytics teams should leverage tools built on these open standards whenever possible in their application development projects to avoid the problems of integrating proprietary formats.

Semantic Analysis

Semantic analysis establishes the meaning of a word or phrase. At this point, ambiguous terms are resolved so that we might know that "a dog of a product" is different from a collie. Dictionaries and patterns of co-occurring words help establish meaning. Semantic analysis helps in categorization and vice versa. It sharpens search results and also improves analysis of content by providing more accurate input to analytics visualizations.

Dictionaries/Lexicons

As we stated previously, domain context is extremely important for the accuracy and, therefore, the usefulness of a content analytics–based application. Industries, organizational structures, professional roles, and even specific task contexts can require highly differentiated, domain-specific term and entity definitions, structures, and relationships. The dictionaries component of the content analytics system enables content application developers to specify their own lists of terms and define their significance in the context of the application. These terms, thesauri, or controlled vocabularies are then available to guide a wide variety of system functions, from initial indexing routines to query response and relevance ranking. Browsing interfaces may rely on thesauri to establish top terms and narrower terms.

Practical Pointers

Leveraging Available Domain Resources

Identify and use available domain resources when possible

There are a surprising number of domain resources available — many free of charge — to content analytics application developers. For example, many federal government agencies publish and maintain taxonomies related to their area of activity. Industry associations also publish very detailed taxonomies or bibliographies; for example, MeSH, the medical taxonomy from the National Library of Medicine that is the metadata guide for the MEDLINE bibliographic database of citations for biomedical literature (life science and health journal articles). WordNet, in multiple languages, provides synonyms and is also used to disambiguate the meanings of words. Structured document standards, such as the XBRL format for public company financial reporting adopted by the SEC, can provide content analytics developers in the financial sector with tools to help deconstruct financial data.

A small group of commercial software firms offer collections of taxonomies covering multiple industries or processes. An enterprise's own organizational structure, line-of-business organization, or employee role definitions from human resources may offer potential resources close at hand for the content analytics application.

Content analytics application developers should be aware of these available resources and think creatively about getting the best value out of them.

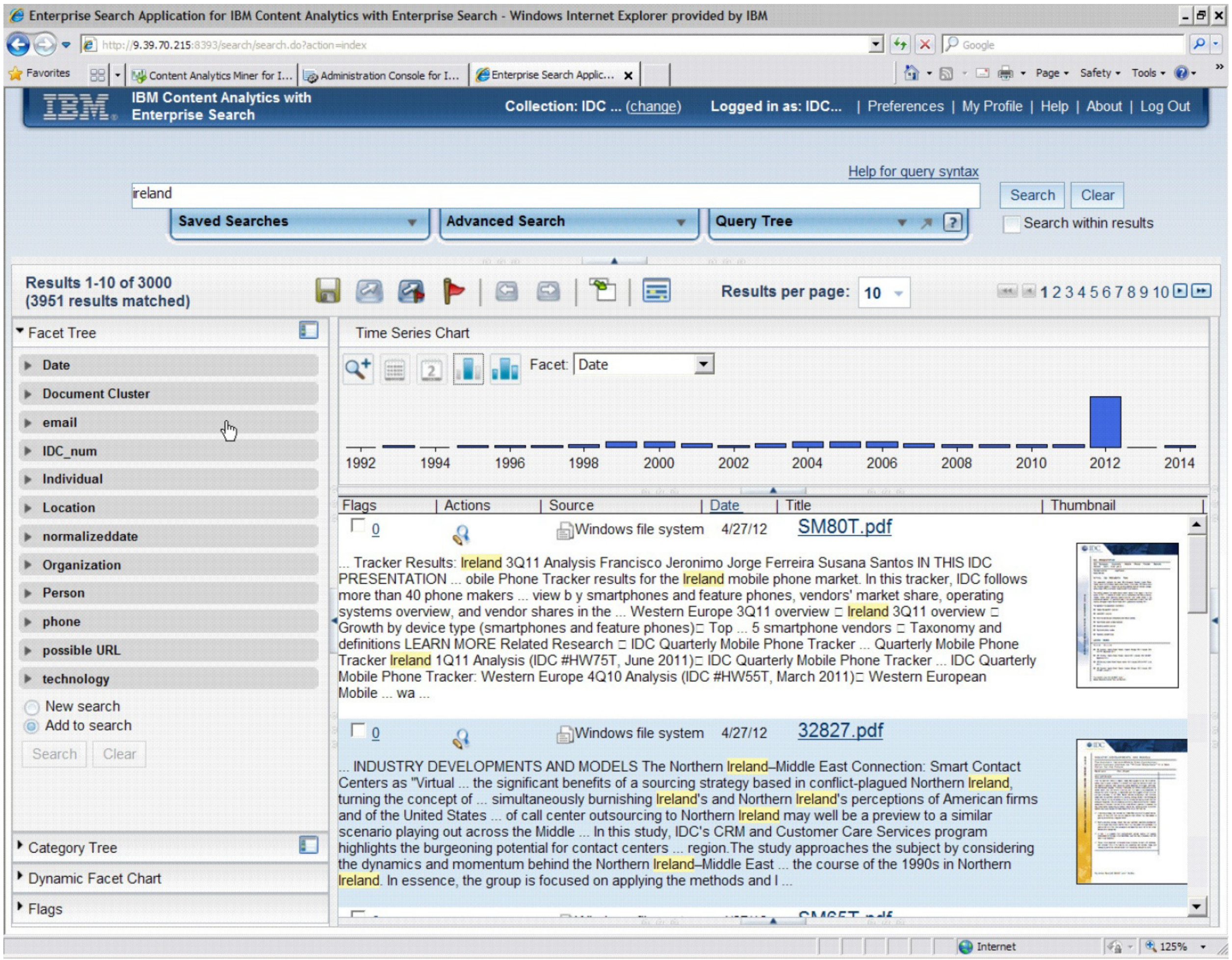
Named Entity Recognition

The named entity recognition component in a content analytics system recognizes the logical and physical definitions of people, places, and things and lays the foundation for analyzing actions and events.

As the screen shot sequence beginning with Figure 3A indicates, a standard search for "ireland" in a repository of IDC research documents renders nearly 4,000 results. But if a user is looking for reports about videogaming by IDC video analyst Greg Ireland, she can leverage named entity recognition in the "Individual" annotator section of the left-frame navigation panel of IBM Content Analytics with Enterprise Search software. A single click dramatically reduces the results list and shows that 196 documents are associated with Greg (see Figure 3B), while a second click, leveraging the "Document Cluster" annotator section, shows that Greg has authored 171 documents germane to videogaming and shows a thumbnail of the most relevant, a PowerPoint conference presentation from an IDC Web Conference in November 2011 (see Figure 3C).

FIGURE 3A

Search Example for "Ireland"



ireland

Results 1-10 of 3000
(3951 results matched)

Results per page: 10

Facet: Date

Flags	Actions	Source	Date	Title
□ 0		Windows file system	4/27/12	SM80T.pdf

... Tracker Results: Ireland 3Q11 Analysis Francisco Jeronimo Jorge Ferreira Susana Santos IN THIS IDC PRESENTATION ... obile Phone Tracker results for the Ireland mobile phone market. In this tracker, IDC follows more than 40 phone makers ... view b y smartphones and feature phones, vendors' market share, operating systems overview, and vendor shares in the ... Western Europe 3Q11 overview □ Ireland 3Q11 overview □ Growth by device type (smartphones and feature phones)□ Top ... 5 smartphone vendors □ Taxonomy and definitions LEARN MORE Related Research □ IDC Quarterly Mobile Phone Tracker ... Quarterly Mobile Phone Tracker Ireland 1Q11 Analysis (IDC #HW75T, June 2011)□ IDC Quarterly Mobile Phone Tracker ... IDC Quarterly Mobile Phone Tracker: Western Europe 4Q10 Analysis (IDC #HW55T, March 2011)□ Western European Mobile ... wa ...

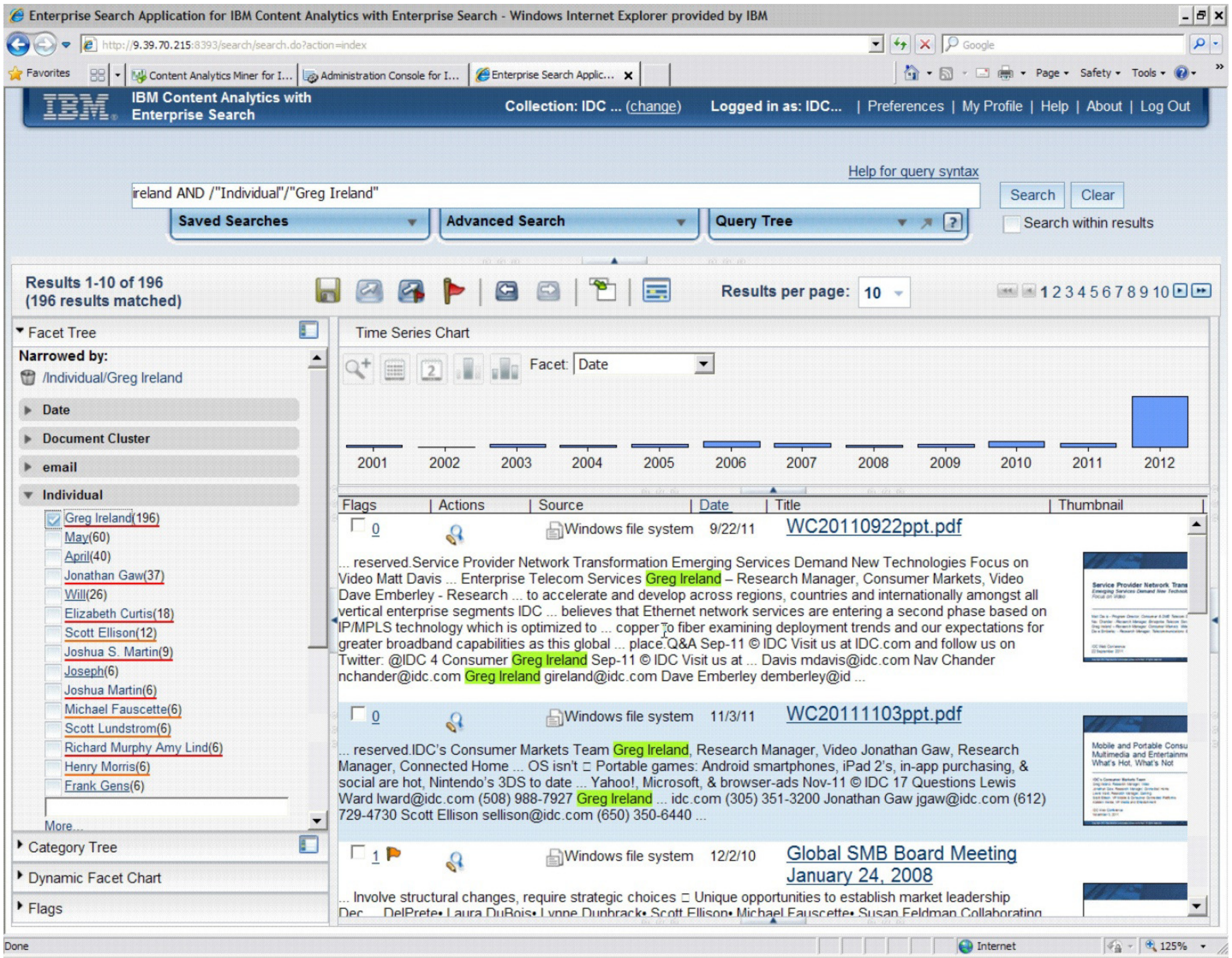
Flags	Actions	Source	Date	Title
□ 0		Windows file system	4/27/12	32827.pdf

... INDUSTRY DEVELOPMENTS AND MODELS The Northern Ireland–Middle East Connection: Smart Contact Centers as "Virtual ... the significant benefits of a sourcing strategy based in conflict-plagued Northern Ireland, turning the concept of ... simultaneously burnishing Ireland's and Northern Ireland's perceptions of American firms and of the United States ... of call center outsourcing to Northern Ireland may well be a preview to a similar scenario playing out across the Middle ... In this study, IDC's CRM and Customer Care Services program highlights the burgeoning potential for contact centers ... region.The study approaches the subject by considering the dynamics and momentum behind the Northern Ireland–Middle East ... the course of the 1990s in Northern Ireland. In essence, the group is focused on applying the methods and I ...

Source: IBM, 2012

FIGURE 3B

Search Example for Individual "Ireland"


Enterprise Search Application for IBM Content Analytics with Enterprise Search - Windows Internet Explorer provided by IBM
http://9.39.70.215:8393/search/search.do?action=index
Google
Favorites
Content Analytics Miner for I...
Administration Console for I...
Enterprise Search Applic...
Page
Safety
Tools
IBM
IBM Content Analytics with Enterprise Search
Collection: IDC ... (change)
Logged in as: IDC...
Preferences | My Profile | Help | About | Log Out
Help for query syntax
Search
Clear
Saved Searches
Advanced Search
Query Tree
Search within results
Results per page: 10
1 2 3 4 5 6 7 8 9 10
Facet Tree
Time Series Chart
Facet: Date
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Category Tree
Dynamic Facet Chart
Flags
Done
Internet
125%

Search query: ireland AND /"Individual"/"Greg Ireland"

Results 1-10 of 196 (196 results matched)

Narrowed by: /Individual/Greg Ireland

- Date
- Document Cluster
- email
- Individual
  - ☑ Greg Ireland(196)
  - □ May(60)
  - □ April(40)
  - □ Jonathan Gaw(37)
  - □ Will(26)
  - □ Elizabeth Curtis(18)
  - □ Scott Ellison(12)
  - □ Joshua S. Martin(9)
  - □ Joseph(6)
  - □ Joshua Martin(6)
  - □ Michael Fauscette(6)
  - □ Scott Lundstrom(6)
  - □ Richard Murphy Amy Lind(6)
  - □ Henry Morris(6)
  - □ Frank Gens(6)
  - More...

Flags	Actions	Source	Date	Title	Thumbnail
□ 0		Windows file system	9/22/11	WC20110922ppt.pdf	
□ 0		Windows file system	11/3/11	WC20111103ppt.pdf	
□ 1		Windows file system	12/2/10	Global SMB Board Meeting January 24, 2008	

WC20110922ppt.pdf: ... reserved.Service Provider Network Transformation Emerging Services Demand New Technologies Focus on Video Matt Davis ... Enterprise Telecom Services Greg Ireland – Research Manager, Consumer Markets, Video Dave Emberley - Research ... to accelerate and develop across regions, countries and internationally amongst all vertical enterprise segments IDC ... believes that Ethernet network services are entering a second phase based on IP/MPLS technology which is optimized to ... copper to fiber examining deployment trends and our expectations for greater broadband capabilities as this global ... place.Q&A Sep-11 © IDC Visit us at IDC.com and follow us on Twitter: @IDC 4 Consumer Greg Ireland Sep-11 © IDC Visit us at ... Davis mdavis@idc.com Nav Chander nchander@idc.com Greg Ireland gireland@idc.com Dave Emberley demberley@id ...

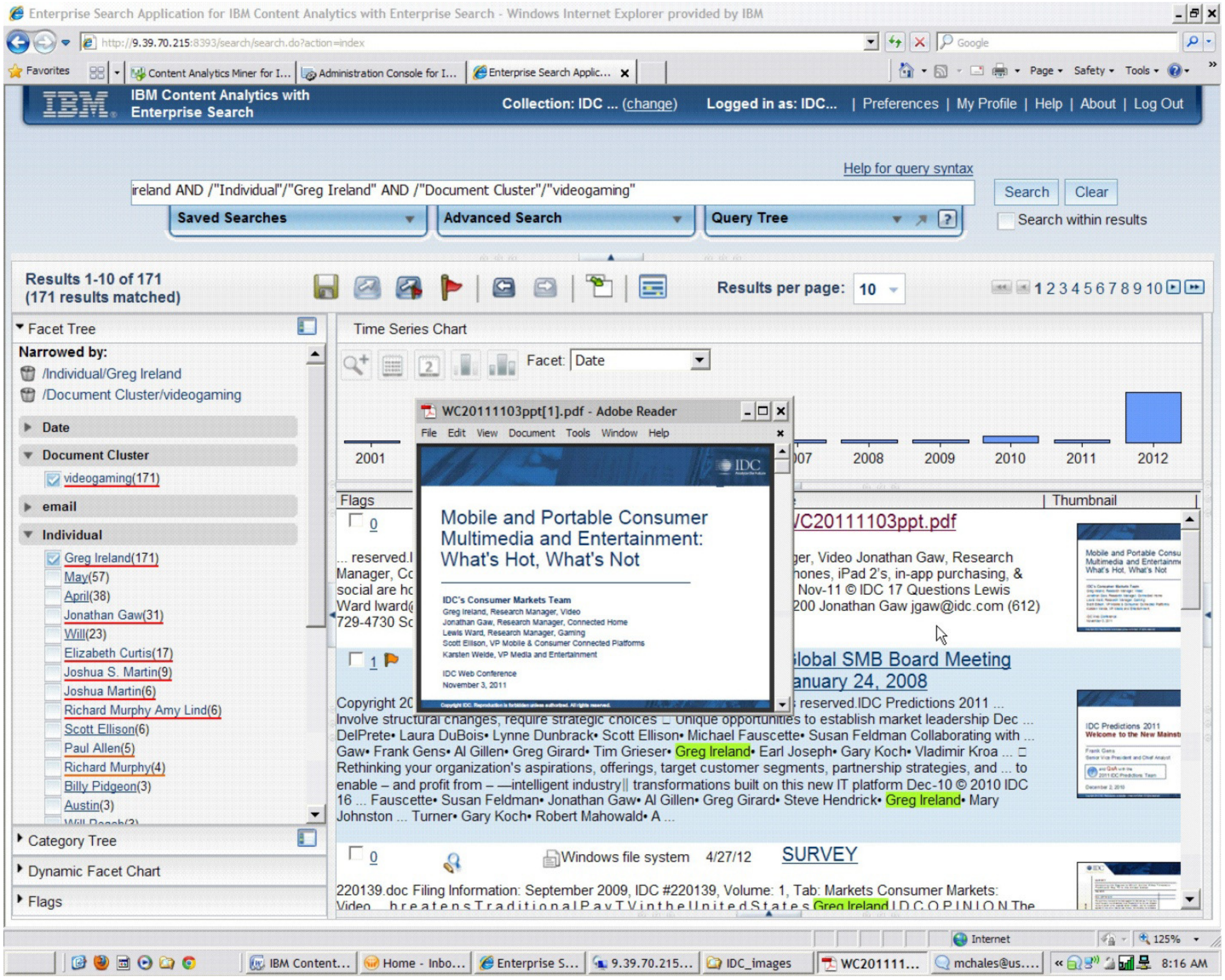
WC20111103ppt.pdf: ... reserved.IDC's Consumer Markets Team Greg Ireland, Research Manager, Video Jonathan Gaw, Research Manager, Connected Home ... OS isn't ⊏ Portable games: Android smartphones, iPad 2's, in-app purchasing, & social are hot, Nintendo's 3DS to date ... Yahoo!, Microsoft, & browser-ads Nov-11 © IDC 17 Questions Lewis Ward lward@idc.com (508) 988-7927 Greg Ireland ... idc.com (305) 351-3200 Jonathan Gaw jgaw@idc.com (612) 729-4730 Scott Ellison sellison@idc.com (650) 350-6440 ...

Global SMB Board Meeting January 24, 2008: ... Involve structural changes, require strategic choices ⊏ Unique opportunities to establish market leadership Dec ... DelPrete• Laura DuBois• Lynne Dunbrack• Scott Ellison• Michael Fauscette• Susan Feldman Collaborating

Source: IBM, 2012

FIGURE 3C

Search Example with Document Preview



ireland AND /"Individual"/"Greg Ireland" AND /"Document Cluster"/"videogaming"

Mobile and Portable Consumer Multimedia and Entertainment: What's Hot, What's Not

IDC's Consumer Markets Team
Greg Ireland, Research Manager, Video
Jonathan Gaw, Research Manager, Connected Home
Lewis Ward, Research Manager, Gaming
Scott Ellison, VP Mobile & Consumer Connected Platforms
Karsten Welde, VP Media and Entertainment

IDC Web Conference
November 3, 2011

/C20111103ppt.pdf

... reserved.I Manager, Cc social are hc Ward lward@ 729-4730 Sc ger, Video Jonathan Gaw, Research hones, iPad 2's, in-app purchasing, & Nov-11 © IDC 17 Questions Lewis 200 Jonathan Gaw jgaw@idc.com (612)

lobal SMB Board Meeting anuary 24, 2008

Copyright 20 reserved.IDC Predictions 2011 ... Involve structural changes, require strategic choices □ Unique opportunities to establish market leadership Dec ... DelPrete• Laura DuBois• Lynne Dunbrack• Scott Ellison• Michael Fauscette• Susan Feldman Collaborating with ... Gaw• Frank Gens• Al Gillen• Greg Girard• Tim Grieser• Greg Ireland• Earl Joseph• Gary Koch• Vladimir Kroa ... □ Rethinking your organization's aspirations, offerings, target customer segments, partnership strategies, and ... to enable – and profit from – —intelligent industry|| transformations built on this new IT platform Dec-10 © 2010 IDC 16 ... Fauscette• Susan Feldman• Jonathan Gaw• Al Gillen• Greg Girard• Steve Hendrick• Greg Ireland• Mary Johnston ... Turner• Gary Koch• Robert Mahowald• A ...

SURVEY

220139.doc Filing Information: September 2009, IDC #220139, Volume: 1, Tab: Markets Consumer Markets: Video ... h r e a t e n s T r a d i t i o n a l P a y T V i n t h e U n i t e d S t a t e s Greg Ireland I D C O P I N I O N The

Source: IBM, 2012

Content Classification

Content classification is a core function of content analytics and encompasses many different kinds of processing routines and potential strategies for deployment. From topic-based clustering, a broad-brush technique often integrated into enterprise search, to advanced machine learning–based automatic taxonomy generation, content classification offers a wide array of options to content analytics application implementers. Some of these options are complementary, such as the use of rules-based predefined taxonomies alongside more flexible, faceted classification approaches. Others require implementers to make a choice between techniques and classes of algorithms.

The ability to surface categories of subject matter within the content is key to allowing users to browse successfully through collections of content. It also offers a drilldown navigation pattern, in which users can "surf" available facets to narrow the universe of content down to just those data elements they are interested in — e.g., beverage/wine; country/France; color/red; grape/pinot noir; region/Burgundy; vintage/2007, etc. This faceted navigation approach, pioneered in commercial use by ecommerce Web sites, is now becoming an industry standard for most information retrieval and analytics applications.

Classification has many uses, not just in user-facing functions but also in back-office or "lights out" routines (e.g., routing only appropriate documents for archiving while retiring others in the context of a records management or eDiscovery system). Content analytics development teams should familiarize themselves with the multiple capabilities and implementation strategies available with clustering and classification software.

Sentiment Extraction

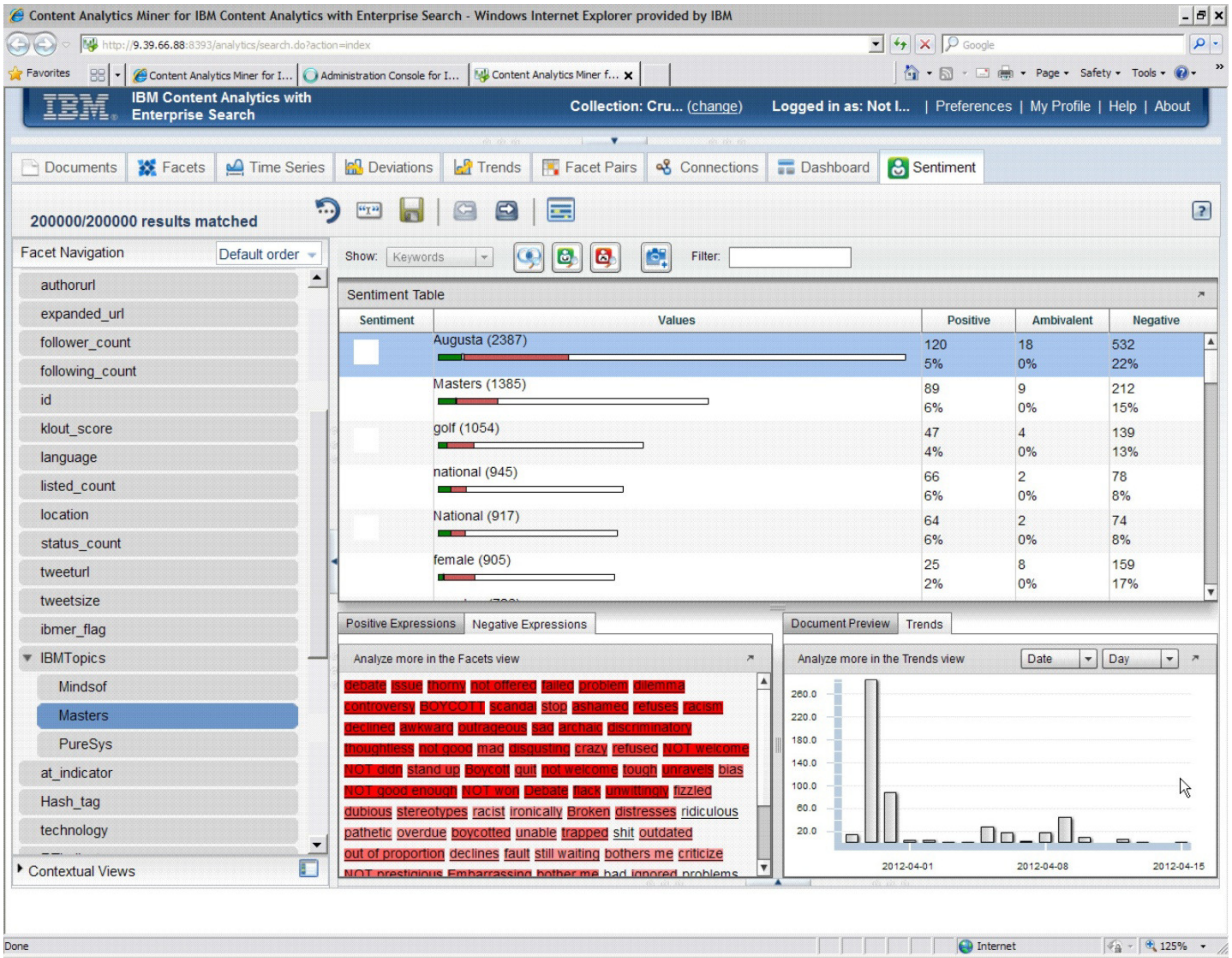
Sentiment extraction is a specialized application of content classification in which the content analytics software is configured to seek out particular terms and phrases that express opinions or evaluations and to categorize them by whether they indicate positive, negative, or neutral evaluations about an entity or event. While sentiment extraction has been in commercial use for some time, its use and its value have increased dramatically with the rise of social media and the explosion of user-generated content in community posts and blogs and in comment streams on company, product, news, or special service Web sites such as Yelp, TripAdvisor, etc.

In addition to leveraging a classification engine, sentiment extraction depends heavily on the quality of the available dictionaries, as evaluative terms often change rapidly in common speech. It is also highly culturally dependent, as value expressions frequently use different terms in different languages or cultures. For example, a Bostonian may call something "wicked good," causing raised eyebrows in Minnesota.

Sentiment can provide insights into information that are not possible with standard content analytics. An example of this can be seen in Figure 4, where tweets are being analyzed around the Masters golf tournament in 2012.

FIGURE 4

Masters Golf Tournament Sentiment Analysis Report



Sentiment Table

Sentiment	Values	Positive	Ambivalent	Negative
	Augusta (2387)	120 5%	18 0%	532 22%
	Masters (1385)	89 6%	9 0%	212 15%
	golf (1054)	47 4%	4 0%	139 13%
	national (945)	66 6%	2 0%	78 8%
	National (917)	64 6%	2 0%	74 8%
	female (905)	25 2%	8 0%	159 17%

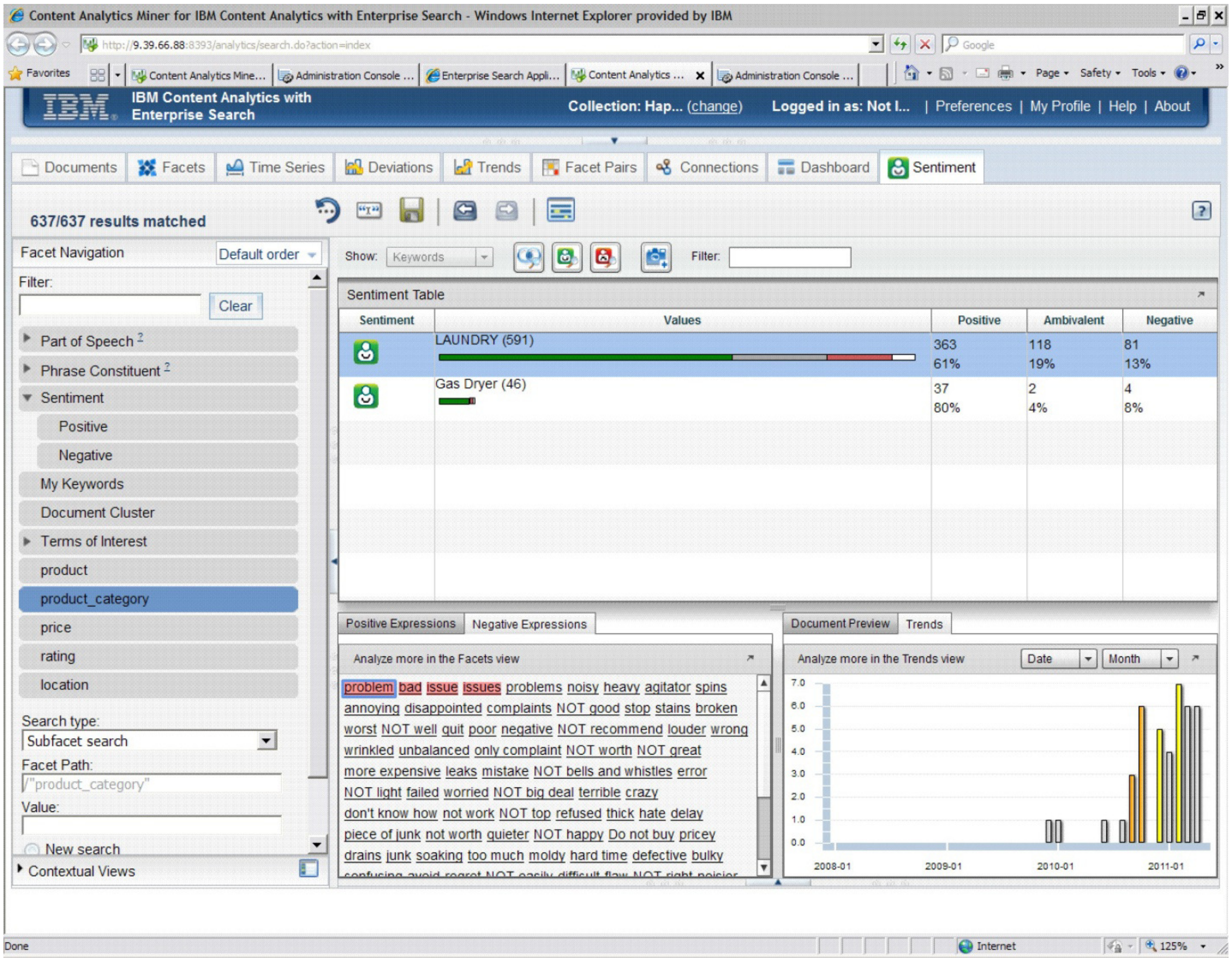
debate issue thorny not offered failed problem dilemma
controversy BOYCOTT scandal stop ashamed refuses racism
declined awkward outrageous sad archaic discriminatory
thoughtless not good mad disgusting crazy refused NOT welcome
NOT didn stand up Boycott quit not welcome tough unravels bias
NOT good enough NOT won Debate flack unwittingly fizzled
dubious stereotypes racist ironically Broken distresses ridiculous
pathetic overdue boycotted unable trapped shit outdated
out of proportion declines fault still waiting bothers me criticize
NOT prestigious Embarrassing bother me bad ignored problems

Source: IBM, 2012

Another example can be seen in Figure 5, which shows sentiment about various appliances. The recommendation for sentiment is that it can be an important part of the content analytics process. Organizations should evaluate whether or not they will need sentiment analysis as part of their overall content analysis strategy and, if so, make sure that the content analytics tool that is used offers sentiment analysis, extraction, and display capabilities.

FIGURE 5

Household Products Sentiment Analysis Report



Sentiment Table

Sentiment	Values	Positive	Ambivalent	Negative
☺	LAUNDRY (591)	363 61%	118 19%	81 13%
☺	Gas Dryer (46)	37 80%	2 4%	4 8%

problem bad issue issues problems noisy heavy agitator spins annoying disappointed complaints NOT good stop stains broken worst NOT well quit poor negative NOT recommend louder wrong wrinkled unbalanced only complaint NOT worth NOT great more expensive leaks mistake NOT bells and whistles error NOT light failed worried NOT big deal terrible crazy don't know how not work NOT top refused thick hate delay piece of junk not worth quieter NOT happy Do not buy pricey drains junk soaking too much moldy hard time defective bulky confusing avoid regret NOT easily difficult flaw NOT right noisier

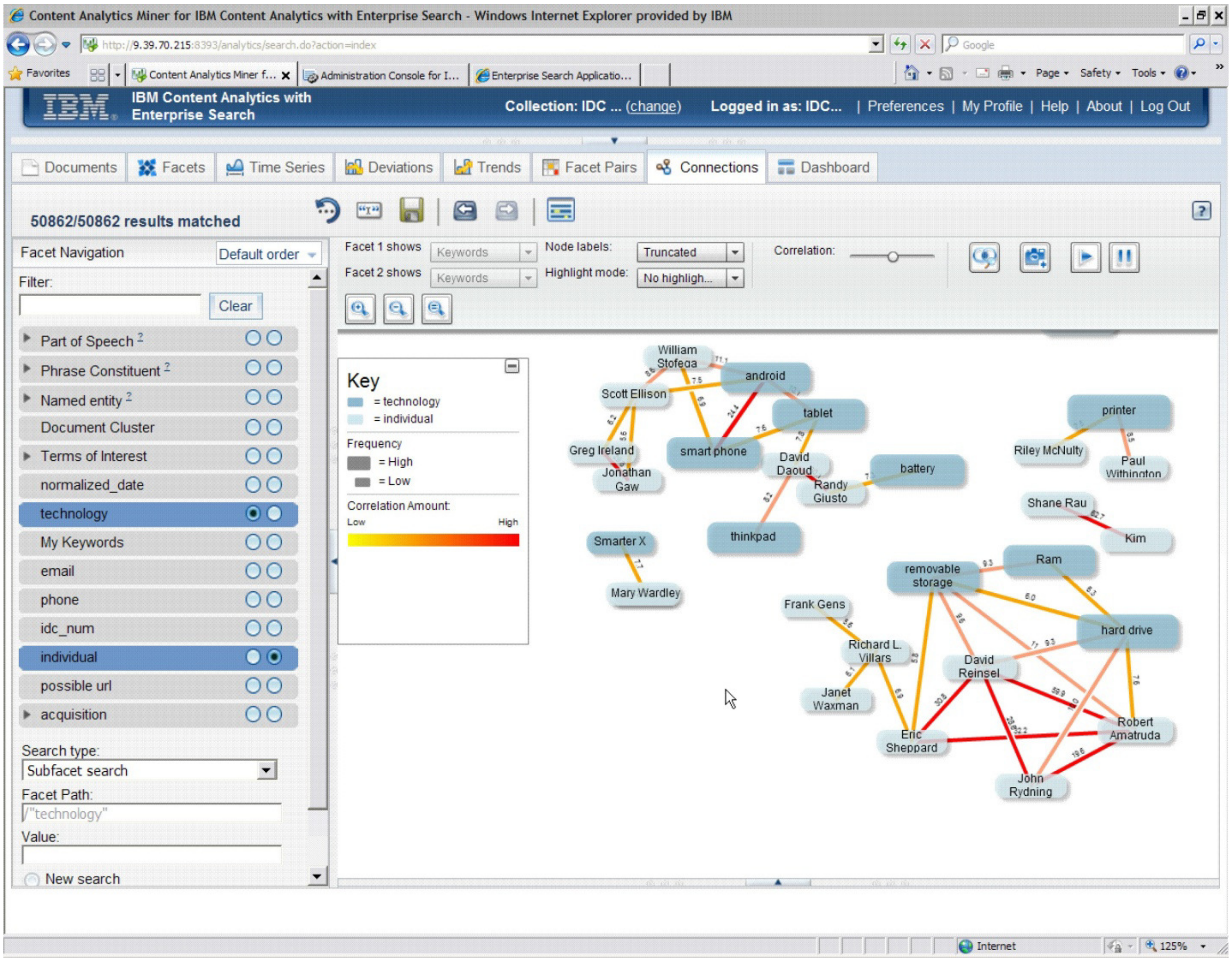
Source: IBM, 2012

Visualization

Visualization is another tool in the content analytics arsenal. Visualization can facilitate information and content navigation and discovery. Correlations between types of entities can show relationships and other types of useful information. For example, the visualization in Figure 6 shows the relationships between the entity types "technology" and "individual" in a collection of articles and reports.

FIGURE 6

Entity Correlation Diagram



Source: IBM, 2012

Practical Pointers

Visualization

Identify visualization needs for the project

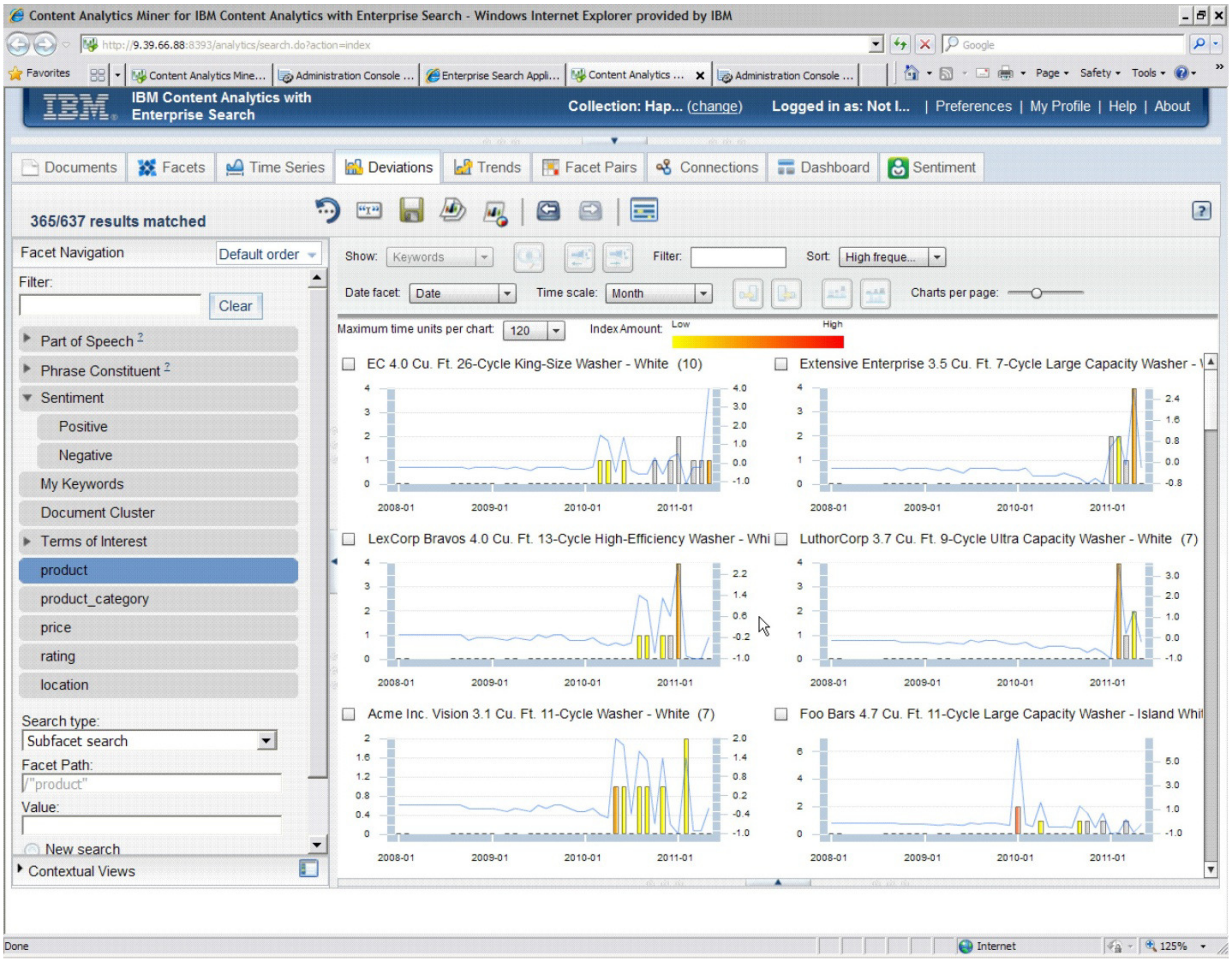
Once information has been extracted or "mined," showing it in a chart or a graphic instead of in text makes it easier for users to understand and identify relationships, trends, and patterns. Simple timelines, bar charts, or pie charts can be used as effectively with mined content elements as they can with data. In fact, the data and the content can be combined in a single graphic for a more complete view.

Content analytics products vary widely in their ability to provide visualization capabilities inside the product. Broadly, content analytics reporting tools and visualizations are parallel in function to the visualization tools available in business intelligence products. They should offer discovery and exploration without extensive customization. Organizations often have their own heritage around visuals for decision making, and content analytics software should be able to feed into these graphics. In addition, the software should provide application programming interfaces (APIs) or gateways to third-party visualization tools. Content analytics development teams should identify their visualization needs, at least at a macro level, early in the project and determine if their requirements will necessitate including a separate visualization tool from the content analytics product or whether the standard visualization in the content analytics product will be sufficient.

There are many types of visualization, and often they can be combined in interesting ways to provide additional information in a limited amount of space. In Figure 7, product mentions for several different types of washers are compared across a time dimension. This type of visualization allows a user to directly review and analyze several different entities on a single screen.

FIGURE 7

Washer Sentiment Comparison Report


Content Analytics Miner for IBM Content Analytics with Enterprise Search - Windows Internet Explorer provided by IBM

http://9.39.66.88:8393/analytics/search.do?action=index

IBM Content Analytics with Enterprise Search

Collection: Hap... (change) Logged in as: Not l...

365/637 results matched

Facet Navigation: Default order

Show: Keywords

Filter:

Sort: High freque...

Date facet: Date

Time scale: Month

Maximum time units per chart: 120

Search type: Subfacet search

Facet Path: /"product"

Value:
Favorites
Content Analytics Mine...
Administration Console ...
Enterprise Search Appli...
Content Analytics ...
Administration Console ...
Page
Safety
Tools
Preferences | My Profile | Help | About
Documents
Facets
Time Series
Deviations
Trends
Facet Pairs
Connections
Dashboard
Sentiment
Clear
Part of Speech 2
Phrase Constituent 2
Sentiment
Positive
Negative
My Keywords
Document Cluster
Terms of Interest
product
product_category
price
rating
location
New search
Contextual Views
Charts per page:
Index Amount
Low
High
EC 4.0 Cu. Ft. 26-Cycle King-Size Washer - White (10)
Extensive Enterprise 3.5 Cu. Ft. 7-Cycle Large Capacity Washer - \
LexCorp Bravos 4.0 Cu. Ft. 13-Cycle High-Efficiency Washer - Whi
LuthorCorp 3.7 Cu. Ft. 9-Cycle Ultra Capacity Washer - White (7)
Acme Inc. Vision 3.1 Cu. Ft. 11-Cycle Washer - White (7)
Foo Bars 4.7 Cu. Ft. 11-Cycle Large Capacity Washer - Island Whit
2008-01
2009-01
2010-01
2011-01
Done
Internet
125%

Source: IBM, 2012

Custom Analytics Components

Most content analytics implementations will need one or more custom annotators to fulfill requirements unique to their application. These might range from a unique approach to processing sentiment to a custom routine for pulling data from call logs or an interface for mining social media activity from a big data cluster.

Each content analytics software product has its own set of supported languages or development wizards to help render custom modules or annotators. There are also some emerging industry standards, which many commercial software packages support. It is important for content analytics teams to evaluate the range of customization capabilities available in their software and to gauge whether those capabilities offer a level of flexibility and ease of use that they feel comfortable with. It is also important for teams to evaluate whether the required technical skill sets and other resources are available on the team, from the software vendor, or from the professional services community. Custom development of content analytics requires a set of skills for which there is currently far more demand than supply.

Practical Pointers

Customization: Scripts Versus Models

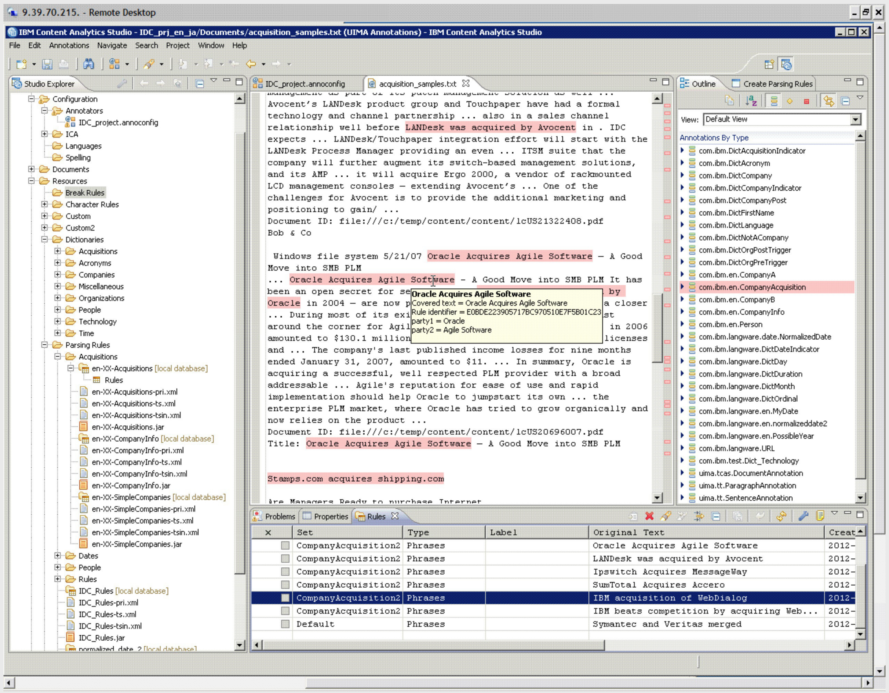
Use advanced modeling tools if available

Content analytics development teams need to demand "out of the box" analytics routines when they are evaluating software. Without such out-of-the-box capabilities, and the ability to mix and match them to achieve customized analytics routines quickly, the teams will have to devote dedicated developers to coding new scripts, many of which will be low-level functionality that have potentially been created hundreds of times in the past, to perform tasks in the analytics pipeline.

State-of-the-art products will offer a wide range of out-of-the box routines as well as a graphical modeling environment that simplifies and streamlines the process of constructing and customizing a content analytics pipeline to meet specific application requirements. The screenshot in Figure 8 represents the kind of interface available in advanced products today.

FIGURE 8

Analytics Modeling Tool Screen Shot


9.39.70.215. - Remote Desktop
IBM Content Analytics Studio - IDC_prj_en_ja/Documents/acquisition_samples.txt (UIMA Annotations) - IBM Content Analytics Studio
File Edit Annotations Navigate Search Project Window Help
Studio Explorer
Configuration
Annotators
IDC_project.annoconfig
ICA
Languages
Spelling
Documents
Resources
Break Rules
Character Rules
Custom
Custom2
Dictionaries
Acquisitions
Acronyms
Companies
Miscellaneous
Organizations
People
Technology
Time
Parsing Rules
Acquisitions
en-XX-Acquisitions [local database]
Rules
en-XX-Acquisitions-pri.xml
en-XX-Acquisitions-ts.xml
en-XX-Acquisitions-tsin.xml
en-XX-Acquisitions.jar
en-XX-CompanyInfo [local database]
en-XX-CompanyInfo-pri.xml
en-XX-CompanyInfo-ts.xml
en-XX-CompanyInfo-tsin.xml
en-XX-CompanyInfo.jar
en-XX-SimpleCompanies [local database]
en-XX-SimpleCompanies-pri.xml
en-XX-SimpleCompanies-ts.xml
en-XX-SimpleCompanies-tsin.xml
en-XX-SimpleCompanies.jar
Dates
People
Rules
IDC_Rules [local database]
IDC_Rules-pri.xml
IDC_Rules-ts.xml
IDC_Rules-tsin.xml
IDC_Rules.jar
normalized_date_2 [local database]
IDC_project.annoconfig
acquisition_samples.txt
Outline
Create Parsing Rules
View: Default View
Annotations By Type
com.ibm.DictAcquisitionIndicator
com.ibm.DictAcronym
com.ibm.DictCompany
com.ibm.DictCompanyIndicator
com.ibm.DictCompanyPost
com.ibm.DictFirstName
com.ibm.DictLanguage
com.ibm.DictNotACompany
com.ibm.DictOrgPostTrigger
com.ibm.DictOrgPreTrigger
com.ibm.en.CompanyA
com.ibm.en.CompanyAcquisition
com.ibm.en.CompanyB
com.ibm.en.CompanyInfo
com.ibm.en.Person
com.ibm.langware.date.NormalizedDate
com.ibm.langware.DictDateIndicator
com.ibm.langware.DictDay
com.ibm.langware.DictDuration
com.ibm.langware.DictMonth
com.ibm.langware.DictOrdinal
com.ibm.langware.en.MyDate
com.ibm.langware.en.normalizeddate2
com.ibm.langware.en.PossibleYear
com.ibm.langware.URL
com.ibm.test.Dict_Technology
uima.tcas.DocumentAnnotation
uima.tt.ParagraphAnnotation
uima.tt.SentenceAnnotation
Problems
Properties
Rules

```
Avocent's LANDesk product group and Touchpaper have had a formal
technology and channel partnership ... also in a sales channel
relationship well before LANDesk was acquired by Avocent in . IDC
expects ... LANDesk/Touchpaper integration effort will start with the
LANDesk Process Manager providing an even ... ITSM suite that the
company will further augment its switch-based management solutions,
and its AMP ... it will acquire Ergo 2000, a vendor of rackmounted
LCD management consoles — extending Avocent's ... One of the
challenges for Avocent is to provide the additional marketing and
positioning to gain/ ...
Document ID: file:///c:/temp/content/content/lcUS21322408.pdf
Bob & Co

 Windows file system 5/21/07 Oracle Acquires Agile Software — A Good
Move into SMB PLM
... Oracle Acquires Agile Software - A Good Move into SMB PLM It has
been an open secret for se                                          by
Oracle in 2004 — are now p                                  a closer
... During most of its exi                                  st
around the corner for Agil                                   in 2006
amounted to $130.1 million                                  licenses
and ... The company's last published income losses for nine months
ended January 31, 2007, amounted to $11. ... In summary, Oracle is
acquiring a successful, well respected PLM provider with a broad
addressable ... Agile's reputation for ease of use and rapid
implementation should help Oracle to jumpstart its own ... the
enterprise PLM market, where Oracle has tried to grow organically and
now relies on the product ...
Document ID: file:///c:/temp/content/content/lcUS20696007.pdf
Title: Oracle Acquires Agile Software — A Good Move into SMB PLM


Stamps.com acquires shipping.com

Are Managers Ready to purchase Internet
```

Oracle Acquires Agile Software
Covered text = Oracle Acquires Agile Software
Rule identifier = E0BDE223905717BC970510E7F5B01C23
party1 = Oracle
party2 = Agile Software

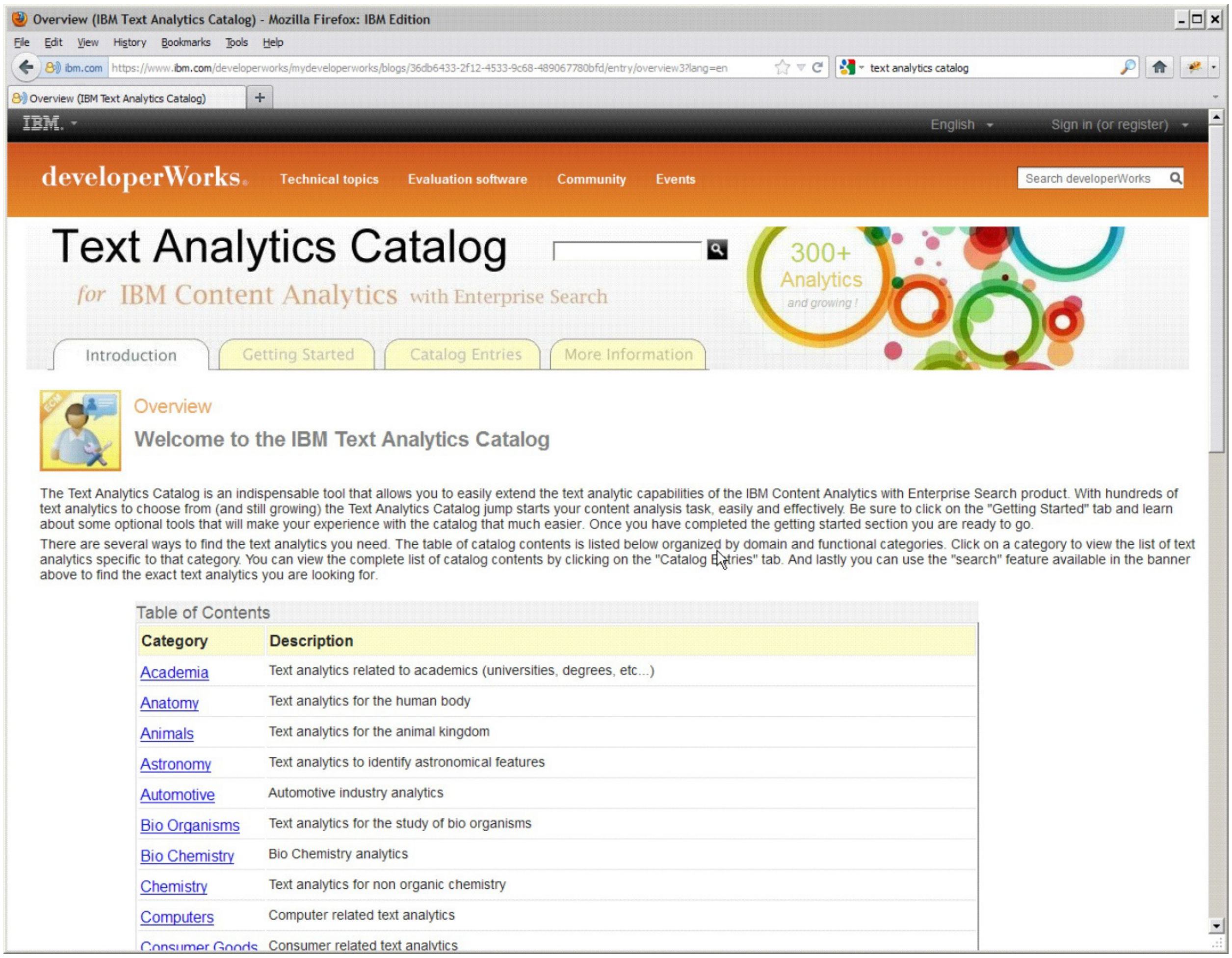
X	Set	Type	Label	Original Text	Creat
□	CompanyAcquisition2	Phrases		Oracle Acquires Agile Software	2012-
□	CompanyAcquisition2	Phrases		LANDesk was acquired by Avocent	2012-
□	CompanyAcquisition2	Phrases		Ipswitch Acquires MessageWay	2012-
□	CompanyAcquisition2	Phrases		SumTotal Acquires Accero	2012-
□	CompanyAcquisition2	Phrases		IBM acquisition of WebDialog	2012-
□	CompanyAcquisition2	Phrases		IBM beats competition by acquiring Web...	2012-
□	Default	Phrases		Symantec and Veritas merged	2012-

Source: IBM, 2012

One method for handling custom development is to use common resources if they are available. In the same manner as there are repositories of taxonomies, dictionaries, and lexicons, a number of content analytics systems have repositories of custom annotators and entity annotators. For example, Figure 9 shows IBM's Text Analytics Catalog for the IBM Content Analytics product. This catalog consists of custom annotators that organizations can develop, share, and use with other organizations.

FIGURE 9

IBM DeveloperWorks Text Analytics Catalog


Overview (IBM Text Analytics Catalog) - Mozilla Firefox: IBM Edition
File Edit View History Bookmarks Tools Help
ibm.com https://www.ibm.com/developerworks/mydeveloperworks/blogs/36db6433-2f12-4533-9c68-489067780bfd/entry/overview3?lang=en
text analytics catalog
Overview (IBM Text Analytics Catalog)
IBM.
English
Sign in (or register)
developerWorks.
Technical topics
Evaluation software
Community
Events
Search developerWorks
300+ Analytics and growing !
Introduction
Getting Started
Catalog Entries
More Information

Text Analytics Catalog

for IBM Content Analytics with Enterprise Search

Overview

Welcome to the IBM Text Analytics Catalog

The Text Analytics Catalog is an indispensable tool that allows you to easily extend the text analytic capabilities of the IBM Content Analytics with Enterprise Search product. With hundreds of text analytics to choose from (and still growing) the Text Analytics Catalog jump starts your content analysis task, easily and effectively. Be sure to click on the "Getting Started" tab and learn about some optional tools that will make your experience with the catalog that much easier. Once you have completed the getting started section you are ready to go.

There are several ways to find the text analytics you need. The table of catalog contents is listed below organized by domain and functional categories. Click on a category to view the list of text analytics specific to that category. You can view the complete list of catalog contents by clicking on the "Catalog Entries" tab. And lastly you can use the "search" feature available in the banner above to find the exact text analytics you are looking for.

Table of Contents

Category	Description
Academia	Text analytics related to academics (universities, degrees, etc...)
Anatomy	Text analytics for the human body
Animals	Text analytics for the animal kingdom
Astronomy	Text analytics to identify astronomical features
Automotive	Automotive industry analytics
Bio Organisms	Text analytics for the study of bio organisms
Bio Chemistry	Bio Chemistry analytics
Chemistry	Text analytics for non organic chemistry
Computers	Computer related text analytics
Consumer Goods	Consumer related text analytics

Source: IBM, 2012

CHALLENGES/OPPORTUNITIES

The Information Advantage

Information finding has become an indispensable part of the online experience: to search for specific facts, people, documents, or images or to help people explore information spaces in order to find trends and unexpected ideas. Content analytics improves both of these processes. It is also part of a larger trend toward pervasive analytics. As organizations attempt to monitor and understand their transactions, their customers, and even their heating and cooling, they can go only so far with standard business intelligence applications. To get the full picture, they must incorporate text and other unstructured formats using content analytics to mine their emails, CRM systems, contracts, and sales and repair records, as well as media blogs, forums, wikis, and tweets outside the organization.

As search itself becomes increasingly central to people's experience of online information — both as enterprise employees and as consumers — the requirement for understandable, relevant, and accurate search results is driving greater adoption of content analytics technologies. The demand from the market spurs innovation among suppliers of content analytics software (e.g., solving the problems of content analytics processing at Web scale).

CONCLUSION

Analytics today is one vital piece of a strategy for outperforming one's peers and competitors. Quick, deeper insights into current trends or previously undiscovered relationships can emerge from content mining. The result is better patient care, new business opportunities, quicker time to market, and preemptive strikes against quickly detected risks.

Business analytics as a core practice in today's enterprise is continuing to grow in importance as more business functions become fully digital. Content analytics software stands at the intersection of the two worlds of structured information and unstructured information, and it plays a central role in the movement toward convergence of these data sources in mainstream analytics.

Organizations cannot afford to ignore the promise of content analytics because the amount of information within the enterprise — already overwhelming — continues to grow at exponential rates. Automated tools for handling this flood of information are the only solution to coping with the increasing demands for wide-ranging and precise business analytics across all the functional areas of the business, from compliance to customer experience to supply chain optimization.

The outperforming enterprises identified in a recent survey by the IBM Institute for Business Value (CEO Study, May 2012, http://www-935.ibm.com/services/us/en/c-suite/ceostudy2012/) were shown to be applying analytics to more aspects of their business and utilizing more sophisticated analyses than their business peer groups. Content analytics provides organizations with a net-new opportunity to address a wide range of critical business problems that feature large amounts of unstructured data or feature a combination of unstructured and structured data. Coming to a more complete understanding of these critical problems more quickly is currently a best practice that allows enterprises to outperform their peers in avoiding significant cost outflows or in gaining the perspective required to take on new business initiatives that increase revenue.