IDC TECHNOLOGY SPOTLIGHT

Unstructured Information: The Revolution in Big Data and Analytics

February 2016

Adapted from *Worldwide Content Analytics, Discovery, and Cognitive Systems Software Market Shares, 2014: Content Analytics and AI Drive Growth* by David Schubmehl, IDC #256823

Sponsored by Sinequa

Content analytics, semantic processing, and machine learning technologies are fundamentally changing the way that information is located, discovered, and utilized by consumers and knowledge workers. These technologies have changed the focus of enterprise search and discovery solutions to become more knowledge and action based, delivering insights, predictions, and recommendations to consumers and knowledge workers on an as-needed basis, as well as creating a new generation of search-based applications. A number of companies are building upon their enterprise search offerings to develop a more complete and robust information access and discovery systems platform.

Introduction

The availability of a wide variety of data spread across many systems and information silos and the technology, skills, and processes to take advantage of it promise to radically change how information is accessed, analyzed, and shared to make better decisions, personalize customer interactions, optimize operations, and innovate. A big part of realizing this promise is dependent on efficient and effective access to unstructured content and the analysis of such content in addition to and in conjunction with structured data. Unstructured content, in particular, is locked in a variety of formats, locations, and applications made up of separate repositories that don't talk to each other.

The value in unstructured content is unique to each organization, and unlocking it is not a trivial task as many organizations face roadblocks because of a lack of appropriate technology and processes. However, organizations that unlock this value have demonstrated an ability to increase revenue, improve productivity, reduce costs, reduce risk, improve compliance — especially around fraud detection — respond to customer or stakeholder needs more quickly and accurately, and bring products to market faster.

According to IDC, unstructured content accounts for 90% of all information, but organizations have generally either significantly underinvested in technology and processes for addressing unstructured content or invested in substandard technology and processes for addressing unstructured content. Yet the ability to find, extract, and utilize the value in unstructured content is more critical than ever. The amount and pace of information that knowledge workers have to deal with on a daily basis are increasing dramatically, and traditional departmental and enterprise search approaches are not keeping up.

The key to solving this dilemma is the use of content analytics, semantic processing, and machine learning technologies, among others, combined with state-of-the-art information access tools to create unified information access solutions that effectively deliver "just in time" insights and value to knowledge workers when they need it most.

Definitions

Content analytics straddles the line between the worlds of unstructured and structured information. 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 databases, repositories of structured data, and collections of unstructured 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 they add this supplementary information as metadata to the search index. Content analytics is embedded into many types of applications. It is often used in enterprise search to improve the overall searchability of information and to provide guideposts that can be used to navigate large amounts of content.

Unified information access platforms provide a single point of access, and they integrate and find relationships in information across multiple heterogeneous sources of information. These platforms have the following characteristics:

- Integrate access to unstructured, semistructured, and structured information
- Combine features of database, business intelligence, and search technologies in a single architecture
- Provide a modular, well-integrated set of tools and services to normalize, index, search, query, present, visualize, analyze, and report information
- Create a single platform for information gathering, analysis, and decision support
- Accommodate quickly changing information through real-time or near-real-time updating and analytics
- Provide a platform for building information-centric applications that support specific industries, tasks, and workflows

Machine learning is the study of computer algorithms that provide computer programs with the ability to learn, discover, predict, and improve automatically using large amounts of data without explicit programming. Although machine learning libraries have been around for decades and have been offered as part of many of the world's statistical packages, the use of machine learning by enterprises hasn't been widespread until recently because of two factors; these algorithms require a lot of data and a lot of compute power. However, many leading technology firms have begun using machine learning tools over the past few years to improve programs in a number of areas, such as image recognition, data categorization, discovery, and information cleansing.

Benefits

Organizations are working with and evaluating a wide range of big data access and analysis options and solutions. However, most of these solutions address only a portion of the challenges that organizations face without a strong foundation in unstructured information handling, access, and analysis. The current and next generations of technologies are based on traditional enterprise search but go well beyond it, offering content analytics, semantic processing, and machine learning to help discover, analyze, and link information in ways that organizations haven't done in the past.

These technologies are forming the foundation of a new information management and access stack for the enterprise. With these new technologies, many leading organizations are adopting the concept of information access "self-service" through the use of automation to link related information and bring it to the knowledge worker directly using the concept of a "logical data warehouse." The "logical data warehouse" is created through the use of software that connects both structured and unstructured repositories via a unified search index and knowledge graph. Organizations are building agile applications based on this "logical data warehouse" that contains curated and enriched knowledge extracted through the use of content analytics and metadata generation.

In many organizations, these systems may replace traditional data warehouses if knowledge workers need quick ad hoc access to large collections of heterogeneous information. If handled correctly, these systems can and will replace, over time, the traditional enterprise search engine and allow corporate developers to create targeted agile applications that solve specific needs, such as an automated customer service app for front-line workers or a tablet-based asset location app for warehouse personnel.

These types of systems can be the foundation for merging data plus content in big data applications, pulling in information from dozens or even hundreds of enterprise applications to present the user with a unified view of a particular problem or challenge.

Trends

The demand for natural language understanding and processing is growing in applications ranging from enterprise search and social media monitoring to cognitive applications and virtual assistant solutions. The strong overall growth rate in a wide range of content analytics represents the beginning of broad adoption of the current generation of unstructured information analytics tools. Artificial intelligence techniques, such as machine learning with neural networks, deep natural language processing and analysis, and a host of other supporting technologies are being combined for use in cognitive systems that can understand questions and directives, hypothesize and formulate possible answers based on available evidence, be trained through the ingestion of vast amounts of content, and ultimately adapt and learn from their mistakes and failures.

In addition, a number of long term trends are driving growth in the space, including:

- The growth of content analytics analyzing unstructured information, adding value via metadata, linkages to structured data, and overall information organization
- The pervasive presence of search as a core interaction pattern in applications serving the enterprise
- The shift to unified information access technologies and the convergence of search with business intelligence
- The emerging development and use of role-based unified information access applications for mobile and tablet-based knowledge workers

Many organizations are also combining traditional data collection and mining with search, content analytics, and sophisticated aggregation technologies to develop new and interesting products. From patent and research offerings to sophisticated localized geographical information, vendors are making available an ever-widening and valuable array of data and data services to customers. New methods for monetizing these data services are also driving the development and use of mobile applications.

Enterprises and organizations should actively consider and plan for these systems within their organizations and/or develop plans for consumer-facing applications that use unstructured information. IDC has found that organizations embracing and making use of both unstructured and structured data and state-of-the-art information access technologies are five times more likely to experience benefits that exceed their expectations than organizations that don't.

Technology vendors should adopt processes and strategies that identify whether or not key data sets and repositories will be required for their applications to be successful. This may also require evaluating the offerings of value-added content companies for other types of data, such as social media data, third-party research, patent information, and competitor information, with the goal of either partnering with or acquiring these companies. Technology vendors should be considering data as a part of their overall solution offering.

Technology vendors need to develop and execute long-range plans for embedding content analytics, information access, and cognitive systems technology within their organizations and as part of their overall business strategies using publicly available data, such as Web data.

In the same way, enterprise content and cognitive system platform vendors will also be offering their APIs as a way for both partners and end-user organizations to quickly and easily add these next-generation capabilities to their applications. Technology vendors should consider these platforms carefully because they may offer distinct advantages to companies that embed these capabilities into their applications.

Considering Sinequa

Sinequa is a big data search and analytics solutions company specializing in the collection, integration, content analysis, and search of large volumes of heterogeneous enterprise data driving customer return on investment and user productivity. Founded in 2002, Sinequa has an extensive foundation of unstructured information access technologies that include scalable search and significant content analytics capabilities in 21 different languages, with full support, including entity extraction, relationship identification, and sentiment analysis for 6 languages, including English, French, and German. Sinequa offers a flexible information collection, access, and analysis architecture where commoditized or high-performance servers can be harnessed to drive the technologies needed to be able to handle unstructured big data application requirements at a reasonable cost or can be used in a cloud-based setting as "platform as a service" (PaaS).

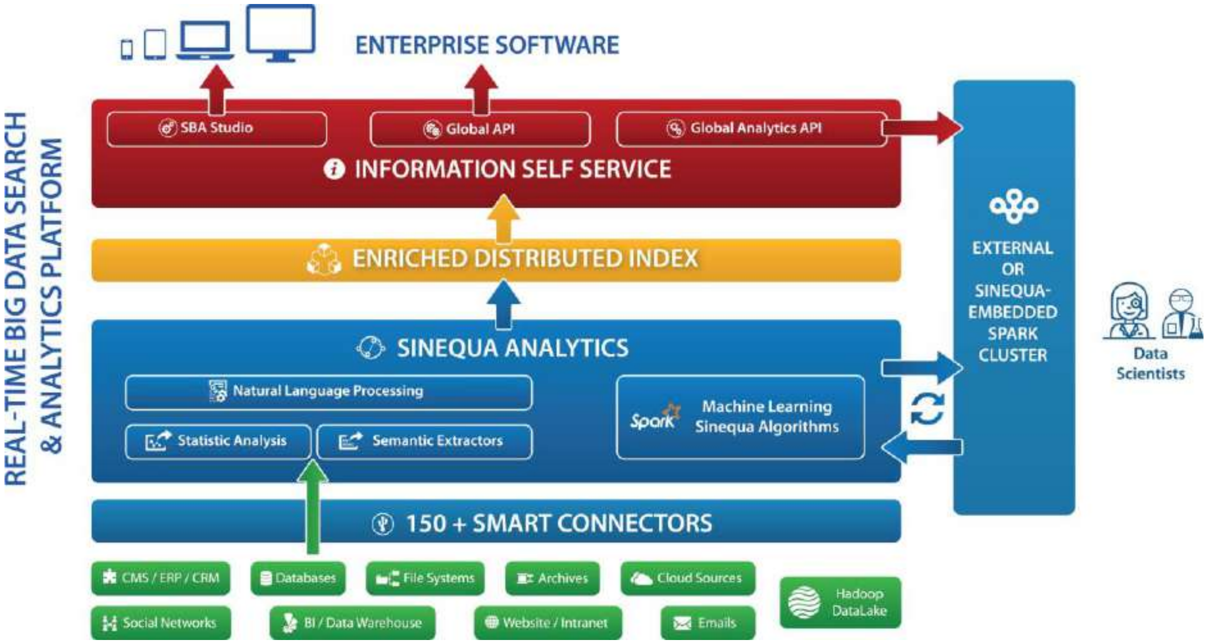
For organizations that need assistance with their unstructured information challenges, Sinequa provides the following:

- It has the technology and capacity to be a player in big data, offering advanced content analytics to find meanings and concepts in very heterogeneous large data sets.
- It offers a unified information access platform for building advanced search-based applications.
- Sinequa's APIs and tools facilitate agile development methodologies, which get search-based applications up and running more quickly and productively than traditional waterfall methodologies.
- It owns the intellectual property for its 150+ connectors and semantic technologies as well as for its search and search architecture. These connectors provide access to big data repositories, cloud systems, and enterprise applications, such as MapR, HDFS, Hive, Microsoft Office 365, Amazon Redshift, and PTC Windchill.
- It has built a number of industry-specific unified information access solutions on its unified information access platform, including solutions for the financial services industry as well as government agencies (e.g., military, police, and intelligence service applications).
- It owns and continues to enhance its natural language processing, content analytics, and application connector intellectual property and is investing in areas such as machine learning and automatic categorization.
- It has millions of users worldwide, including in marquee accounts such as Airbus, AstraZeneca, Biogen, BMS, UCB, Nasdaq, Siemens, Credit Agricole, the French Ministry of Defense, Mercer, Solvay, and Atos.

Sinequa offers a broad-based information collection, access, and analysis platform including search, content analytics, semantic understanding, and auto-categorization technologies that can be used to develop targeted information access solutions ranging from accelerating clinical trial research to providing a 360-degree customer review for an organization's sales and support team, as outlined in Figure 1.

FIGURE 1

Sinequa Architecture



Source: Sinequa, 2016

Challenges

While Sinequa offers an extremely strong platform for unstructured information handling, the company is currently accelerating product plans for its upcoming release around the adoption and use of machine learning and artificial intelligence as requirements for more "smart" applications emerge. Content analytics and cognitive systems technologies have changed the focus of enterprise search and discovery solutions to become more knowledge and action based, delivering insights, predictions, and recommendations to consumers and knowledge workers on an as-needed basis. A number of companies are building upon their enterprise search offerings to develop a more complete and robust cognitive systems platform.

The fact that Sinequa is actively adding cognitive systems capabilities such as machine learning, hypothesis generation, and predictive analytics to its offerings in order to compete more effectively with the pure-play cognitive systems vendors is an interesting/promising move. Because cognitive systems are highly reliant on unstructured information analysis and manipulation, companies such as Sinequa that offer strong capabilities in these areas should be able to build very competitive cognitive systems offerings on these assets.

Conclusion

Enterprises and organizations should actively consider and plan for role-based unified information access systems within their organizations and/or develop plans for consumer-facing applications that use unstructured information. The business benefits of harnessing unstructured information are too great to ignore, and previous approaches of using departmental or traditional enterprise search have been largely unsuccessful. However, search applications that can effectively utilize content analytics, semantic understanding, and machine learning can provide organizations and their users with timely relevant information that will impact the bottom line and significantly increase return on investment.

We look for the following long-term trends to support continued growth in the unstructured information access and analysis market:

- The growth of content analytics analyzing unstructured information, adding value via metadata, linkages to structured data, and overall information organization
- Data outside of the organization in cloud-based applications and the trends toward cloud-based operating
- The increased usage of voice recognition, machine translation, and natural language processing technologies as part of a comprehensive suite of solution development tools
- The shift to unified information access technologies and the convergence of search with business intelligence, which fuels the growth of content analytics
- The emerging development and use of role-based unified information access applications for mobile and tablet-based audiences (These applications become "virtual assistants," providing recommendations and answers and relying on natural language understanding and processing for human-computer interaction.)

IDC believes that the overall market for semantically enabled information access and analysis systems will continue to grow at a significant rate, and to the extent that Sinequa can address the challenges described in this paper, the company has a significant opportunity for success.

ABOUT THIS PUBLICATION

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

For more information on IDC, visit www.idc.com. For more information on IDC Custom Solutions, visit http://www.idc.com/prodserv/custom_solutions/index.jsp.

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