



*Which of the following current topics will significantly change the market in the next year? And what is the impact? (Business Intelligence/Analytics, Customer Engagement, Mobile, Security, or Social)*

Software companies whose applications manage unstructured content know that Big Data is very real, and the need to manage and effectively use the data provides an opportunity for all of us. What's less known is that there is a game-changing technology that's has been proven to address the challenges of managing the ever-growing collections of unstructured Big Data. The solution is designed to be integrated into existing solutions and has demonstrated its ability to manage large amounts of unstructured data in multiple markets.

CAAT from Content Analyst Company was built from the ground up to be integrated into existing solutions. CAAT uses patented mathematical algorithms to understand the conceptual meaning and relationships of terms and documents in any size collection. The technology is delivered as a partner embeddable platform and has been proven to handle the scale of today's document collections in the US Intelligence Community and the highly-regulated world of eDiscovery among other markets.

One of CAAT's popular capabilities is enabling solution providers to synthetically transfer human taxonomy knowledge across an entire organization's electronic documents and emails through our innovative example based auto categorization approach. Example-based auto categorization is faster, easier and far more accurate than traditional lexicon-based taxonomy alternatives. The application of example-based auto categorization has proven that it's no longer necessary to manually create – and constantly maintain – word-based taxonomies and complex rules in order to precisely and accurately classify large volumes of unstructured big data and improve "findability" of information.

Content Analyst Company partners with dozens of software companies and systems integrators who use the technology in their solutions to solve a wide-range of business problems. Partners in the areas of legal e-discovery, patent research, social media monitoring and U.S. Intelligence have demonstrated the fast, easy, and repeatable way to pinpoint only the most important documents and emails among collections spanning tens and hundreds of millions of files and messages. Two other key advantages to using CAAT are it is an in-memory technology delivering extreme performance and it is language agnostic; meaning it can work with any language because it learns from the data, not predefined word list or rules. The example-based auto-categorization capability powers predictive coding applications in eDiscovery that have been accepted by the courts as a sound, defensible method for efficiently handling the growing volumes of data in today's cases.

Software companies and content providers have come to realize that semantic advanced analytics technology can understand the 'meaning' of unstructured documents, and is a key in the drive to taming the unstructured content of Big Data. Software companies that have built enterprise content management, cloud and storage management, and archiving applications; to online content publishing or DaaS (Data as a Service) – all face the challenges of managing vast amounts of unstructured content. By taking a small number of documents as examples, and using example-based auto-categorization to say "go find more like these," the potential impact on taming the issues associated with Big Data is actually manageable.

To put a finer point on it, here are some examples of how example-based auto categorization can drive value for enterprises struggling to reduce the burden of big data while increasing the benefit of big data.

- Records Management and Content Management – Example-based auto categorization can be used to enable greater precision in determining which category or categories a document or record should be in. The example-based approach makes it easy to add and maintain categories so manual methods are no longer needed. With CAAT's conceptual understanding of the category it understands conceptually which documents are matches, even if the words are not used in any of the examples. This is why code words, abbreviations, or new terms are not missed or mis-categorized.
- Storage and Archiving – Clear rules exist for "Records," but the majority of unstructured content is unclassified. In many cases, the documents (for example emails) quickly have no business value and have past the required retention requirements. . By quickly defining categories for both "business valued" documents and "junk" (holiday emails, sport fantasy, old marketing documents, etc.), good information management practices can be applied and the defensible deletion is possible.
- Content Publishing and WCM – "Findability" is money for online information providers. When a new hot topic (i.e., a category) emerges, how quickly can you define the rules and words in your taxonomy so your customers can find this topic in their navigation and the related

documents? By augmenting the workflow with an example-based auto-categorization step, content improves its “findability” and relevance, and makes it easy to create, maintain, and add new categories quickly, in any language; equaling revenue and satisfied customers.

- Enterprise Collaboration – Concept-based auto categorization makes documents much easier to find, dramatically improving collaboration, sharing and syndication of your valuable content. With internal research assets and intellectual property that can be leveraged elsewhere in the enterprise, or content generated for external consumption, auto categorization dramatically improves the ability of users to consume and properly apply these information assets.

- Business Intelligence and Analytics – There are numerous players in the structured analytics and visualization space, but how do you create “informational” structure from “unstructured” text? Example-based auto categorization provides rapid classification of unstructured content based on actual concepts contained within.

- Social Media Monitoring – With the volume and velocity of social media content exploding, new terms and meanings are emerging and changing daily, making it nearly impossible to keep up with terminology and common threads of the users’ messages/postings. With conceptual understanding of the information and how new terms are being used just like other known terms, companies can track the real time trends, versus missing key insights because their system did not understand the term, being used by customers.

Despite the hype around big data, few will disagree that it poses challenges and benefits if managed properly, and fewer still will disagree that it’s going away anytime soon. Relying on manual taxonomies is simply not practical, as the volume, velocity and variety of content comprising big data accelerates virtually at the speed of thought.

Concept-based auto categorization has proven itself as a highly effective, extremely fast and incredibly precise approach. The possibilities are endless for applying this technology to address the major obstacles big data poses, while simultaneously harvesting the broad benefits big data stands to offer.

CAAT offers many other advanced analytics capabilities in the same partner embeddable platform. We focused on one capability to help describe what is possible in a few use cases. We, the computer industry, have created the Big Data opportunity by solving the limitations of computers and network bandwidth of a decade ago. Now sending and receiving digital content is the optimal way to deliver information, hence we now face the extreme volume, velocity and variety of information flowing over the high speed networks. We now need to augment or replace applications to keep up with the scale, performance and flexibility requirements of “Big Data”. Advanced Analytics is one approach that now can be applied in multiple areas to make Big Data a company asset versus a burden.

*This interview was published in [SIIA's Vision from the Top](#) , a Software Division publication released at [All About the Cloud 2013](#)*