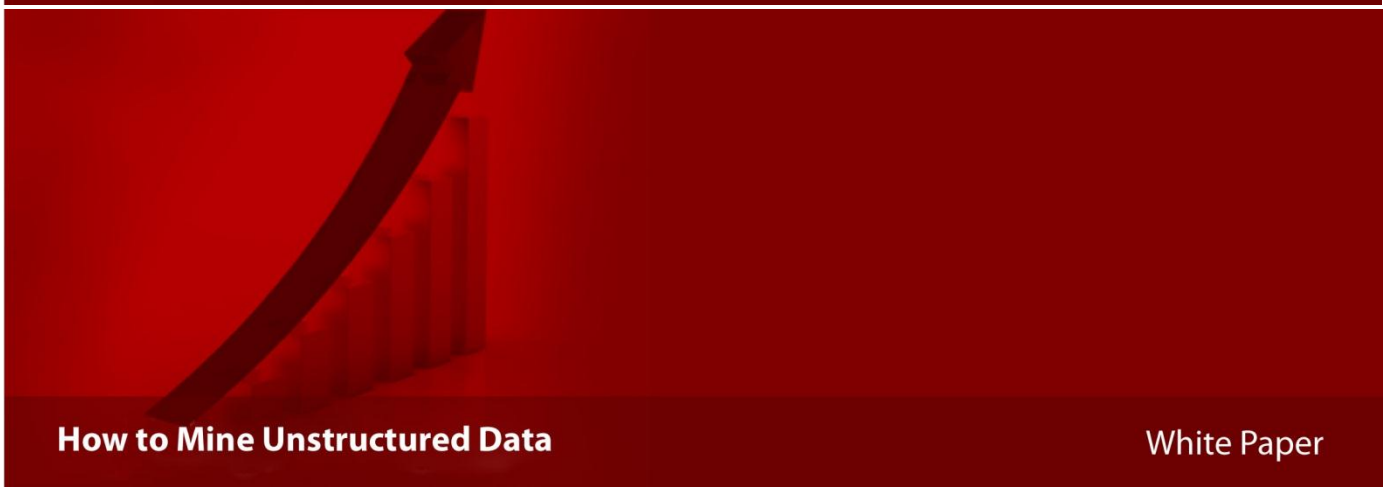

How to Mine Unstructured Data White Paper

July, 2012
www.mu-sigma.com

How to Mine Unstructured Data

In 2003, Information Management published a series of articles, starting with a missive titled "The Problem with Unstructured Data." In what seems to be a generation ago, the big problem identified was many tools and techniques' inability to make sense of unstructured data. Not surprisingly, much has changed in the last decade.

On one side, the explosion in data has accelerated, thanks to a combination of innovation (YouTube started in 2005 and today, users upload 35 hours of video every minute), falling costs of data storage and transmission (cheap Web access over the phone lines has brought millions more to the Internet) and, perhaps most significantly, the emergence of mobile technologies (mobile data traffic alone in 2010 was three times the entire size of the global Internet in 2000). Most of this growth has been in the form of unstructured data, such as video files, audio clips, blogs and posts and conversations on social networking platforms.

Spurred in part by this explosion in data, there has also been an impressive influx of technological innovations aimed at making sense of unstructured data. As these technologies go mainstream, it has becoming increasingly evident that unstructured data presents a new set of opportunities. Some of the possibilities:

- Up until now, customer care centers have focused on traditional metrics, such as call volumes, call lengths and other operational aspects, in order to better manage and improve call handling capabilities. But with sophisticated natural language processing algorithms, call transcripts can now be mined to generate important insights into what the customers are thinking about products and services.

- Traditionally, customer surveys have been the only means for companies to reach out and understand perceptions about their products and services. As any statistician worth his or her salt knows, any sample runs the risk of incorrectly representing the overall population, in addition to the fact that surveys rarely capture the customer experience at the moment of consumption. Enter the world of blogs, tweets and social networking; these tools have given people a channel to express their perceptions about products and services in near real time. Social media data is proving to be an invaluable source of information for companies to understand customer sentiment as a feedback loop. In addition, it is a well-known fact that social networks are dominated by a small set of key influencers. For companies trying to improve brand perception, focusing on these key influencers is turning out to be a far more effective than traditional advertising. What used to be a mass of unstructured data, in the form of conversations and comments, is turning out to be a highly effective window into the minds of customers.

Where Do We Go from Here?

It is no longer an option for companies to ignore the vast amounts of unstructured data that is being accumulated, within and outside their firewalls. And as technologies and data management techniques have evolved, mining for gold in unstructured data is no longer an "occult science"; companies can approach it in a structured manner.

Create a platform (infrastructure and algorithms) to process large quantities of data. Unstructured data typically appears along with its sibling, big data. It is very important to choose data sources with care. Companies typically

identify and store only data that solves business problems. However, there are benefits to conducting exploratory analysis of unstructured data without a business purpose. For example, Twitter users generate 12 terabytes of data every day. Given this scale, it is important to explore new data sources.

To handle unstructured data, many companies try to use modified version of RDBMS. RDBMS technology is more than 30 years old and is not sufficient to handle neither the complexity nor the scale of unstructured data. Big data needs scale: scalable storage, scalable computing, etc. The current favorite with most companies is the open source Hadoop, one of many available tools. A host of vendorshave created integrated offerings to process large amounts of data.

Considering the fact that it is very expensive to create scale, building platforms in the cloud should be seriously considered. Some companies have already announced support for Hadoop on their cloud offerings. There is also a lot of focus on in-memory analytics that can support big data.

Establish a machine-processable structure. Because the data is unstructured, it is important to apply some kind of structure so analysis can take place. For example, a company collecting tweets may want to capture additional information about the tweets, such as language, geography or user ID, to help in further processing. Since we are dealing with unstructured data, the structure imposed on it may change drastically, based on initial results and new data. It is important to note that the data may never be complete and organizations have to make tradeoffs in accuracy. Establishing a structure can be achieved via algorithms or manually, through heuristics. Algorithms reduce the accuracy but achieve scale. Humans, on the other hand, increase accuracy but reduce scale. It is important to have a data model that is flexible, as well as capable of constantly evolving.

Create smaller representative data sets. Not every stakeholder will have the capability to work on unstructured data. It is important to democratize the results by creating smaller structured data sets that address specific business needs. These data sets can then be queried upon using the regular methods of data analysis. Companies should also utilize the benefits of sampling. Many analyses can be performed to a reasonable level of accuracy, with much shorter turnaround times. The random sample cuts data size and cost, by an order of magnitude, and still gives reasonably accurate results. Many times, the analyses performed on smaller data sets could lead to insights, which could be used to deep-dive using the complete data.

Develop/acquire algorithms. There are several kinds of approaches to process unstructured data. Companies can use heuristics, as well as machine learning techniques, to process data. For text mining, natural language processing, combined with neural networks, allows companies to capture the sentiment of social media feeds. Support vector machines allow classification of graphics that enable companies to identify whether visitors are customers or store associates. Techniques, like Bayesian Networks, can help discover patterns across multiple dimensions, allowing organizations to learn by mining existing data. There are several algorithms that are freely available through open source communities. However, it is important to note that you will almost never get it right the first time. Companies should also invest in good visualization techniques. Given the volume of data, visualization is a key factor in ensuring better consumption of insight from the data; often, visualization becomes as important as the analysis.

Develop policies to constantly review and purge data. Figuring out what data to delete is almost as important as acquiring data. There are several startups that have created a business model out of data destruction. Companies need to brutally prioritize what data they should retain. They should develop and implement strict data purge and archival policies. The focus should be on increasing ROI and efficiency. Learning from the most recent information should be a priority, so as to not have historical biases.

Gaining insight from unstructured data is difficult but not impossible. Recent interest and developments in the field have made it easier for companies to work on this data and generate insight. Companies should start exploring unstructured data as the next step in information management.

Krishna Rupanagunta is a Delivery Unit Head for Mu Sigma. In this role he is responsible for client delivery, team development and providing thought leadership across projects. His background is in Business Consulting, servicing Fortune 500 clients across multiple industries with specific focus on Supply Chain Optimization. Prior to joining Mu Sigma, he was part of a non-profit that helped the Indian government in the conceptualization and design of the largest citizen identity project ever attempted in the world. He has a deep interest in Behavioral Economics and actively contributes to the Mu Sigma blog.

David Zakkam is a Senior Delivery Manager for Mu Sigma. He has 10 years of experience working in the analytics industry. His current focus areas include Rapid Impact Analytics, Search Engine Monetization, Big Data and CRM. He has an MBA from the Indian Institute of Management, Calcutta and received an Engineering degree from Indian Institute of Technology, Delhi.

Harsha Rao is a Senior Manager at Mu Sigma where he oversees delivery for clients in the financial services and technology space. His current interests lie in the space of Big Data as well as Risk Management. Prior to Mu Sigma, Harsha was at a leading U.S. bank where he focused on credit risk management. He earned his Ph.D. in Statistics from North Carolina State University.