Fundamental Mind Shifts for the Future of Data Analytics

James Purchase

Abstract

Data analytics requires a fundamental mind shift. Rather than continuing to use myopic information (such as concluded events, past transactions, and in-the-moment conversations), data analytics is entering a new stage—one that strings together historic data with current events, nuances, and (until now) largely unstructured pieces of data to create The Big Picture. With this new school of thought about measuring these strings of data—more colloquially referred to as measuring sentiment—how do companies that have been built on structured data shift their mindset to read the conversational context of data? How do they switch gears to combine history, linguistics, and algorithms to interpret human intention?

This article looks into the future of data analytics to explore how accurate measures of sentiment are, dive deeper into the granular subtleties of language for non-sentiment profiling, and build a timeline beyond sentiment that moves into actual profiling, context space influencing, and what will be the Holy Grail of analytics: dynamic search.

James Purchase is the vice president of product management at Attensity, a provider of corporate insight solutions based on proprietary data contextualization. JPurchase@attensity.com

Introduction

At the South by Southwest Interactive (SXSWi) conference in 2007, Twitter—then a year-old microblogging social network with a handful of users—indelibly altered the way we began listening to conversations online. Even though it replaced nothing, nor made any improvements to an existing way to communicate, by 2010 tweets were being transmitted from the International Space Station, and as of 2013 we saw a peak of 143,199 tweets per second. Today, Twitter is the third most-visited site on the Web, with 645,750,00 users pushing 2.1 billion searches each day.

Despite these mind-boggling statistics, consider this: tweets—on their own—are totally meaningless. To be

fair, Facebook and Google+ posts, Amazon reviews, YouTube, Reddit, Pinterest, blogs, forums, and review sites are, too. At their core, they are all standalone declarations that make up a world of online nonsense.

As a first step in trying to understand and capitalize on online sentiment, descriptive analytics kicked off a numbers game of collecting and analyzing page views.

This data sits in silos of concluded events, past transactions, and in-the-moment conversations that are incapable of transcending to something collective or (what would be even more useful) something closer to *profiling*. Strip away the silos to draw lines between one-off posts and you begin to see a business intelligence scenario of human relationships. Suddenly, trends appear, opportunities can be anticipated, and business decisions can be refined and defined. All this leads us to the next generation of text analytics: beyond sentiment.

Unstructured Data

Jeff Bezos started Amazon in 1995, when just 16 million people used the Internet, but he exploited a startling statistic: the Internet was growing at a rate of 2,300 percent annually. A year later, users numbered 36 million, a figure that would continue to grow at a furious rate. Today, more than 1.7 billion people—or almost one of every four humans on the planet—are online.

Bezos understood two things. One was how the Internet made it possible to banish geography, enabling anyone with a connection and a computer to browse a seemingly limitless universe of goods and opinions with a precision never previously known and then buy items directly from the comfort of home. The second was how—on the other side of the coin—the Internet allowed merchants to gather vast amounts of personal information about individual customers.

When online search analysis first appeared, it relied exclusively on structured data to monitor online transactions and activities. This was information that lived by a set of rules, meta tags, and markups, allowing very specific data to be easily captured, read, and understood by browsers and search bots—but it revealed little about who was talking and what they intended to do.

When video, social media, and aggregators began to appear online, "sentiment" crept into the equation, and marketers took note. This data was less structured than meta tags, expressed interest and viewpoints (both positive and negative), and was more difficult for marketers to follow, capture, or decipher. However, it was a gold mine of relationship and trending intelligence. Although new channels such as Facebook, Twitter, YouTube, Flickr, and Redditt permitted more human interaction and intention to shine through, they initially also represented a new type of information only accessible to trained (and expensive) analysts and data scientists.

Enterprises recognized that their customers were actively talking about their products and services on millions of sites across the social Web. Buried in these conversations were valuable insights that would begin to have a significant impact on business, and businesses wanted to understand these conversations.

The first analytics model to appear on the scene was descriptive. Descriptive analytics is the discipline of summarizing large sets of data in smaller sets with the goal of (ultimately) discovering patterns and useful nuggets of information. As a first step in trying to understand and capitalize on online sentiment, descriptive analytics kicked off a numbers game of collecting and analyzing page views, likes, posts and re-posts, links, mentions, tweets, check-ins, fans, and followers.

Soon, people in departments spanning marketing to brand management to product development to human resources began paying attention to descriptive analytics. Influencers and candidates were ranked by their number

of Twitter followers, blog visitors, and Facebook friends, yet these metrics offered little besides meaningless lists and numbers of past transactions and actions.

The evolutionary next step was predictive analytics—the application of machine learning techniques (modeling and statistics) to examine past and current data. Predictive analytics is an imprecise term because it implies predicting the future, but what the discipline truly offered was predominantly a method for forecasting possibilities or an estimation—much like deducing that the combinations of an achy knee plus a white ring around the moon plus a cloudy sky must mean rain ahead—although it doesn't guarantee it. In Netflix's case, because a viewer chose the Sylvester Stallone movie *Rambo*, predictive analytics suggests that the customer will likely be interested in *Demolition Man* or *Rocky*.

Predictive analytics is an imprecise term because it implies predicting the future, but what the discipline truly offered was predominantly a method for forecasting possibilities or an estimate.

Today, organizational needs are moving beyond traditional descriptive and predictive analytics to advanced analytics, which rolls predictive modeling, clustering, affinity analysis, and optimization into a more robust, prescriptive analysis formula. This is a move to offer real-time data observation and data distillation, with the shift occurring from systems that primarily aggregate and compute structured data toward analytic systems that correlate and relate structured and unstructured data. The objective is a system able to reason, learn, and deliver prescriptive recommendations.

With the growth of cloud, mobile, and social media, the way marketers analyze data in the pursuit of spotting trends, advantages, and threats to develop campaigns and create ROI has drastically changed in recent years. Although it is faster and easier now to get (almost) immediate data mining results to instantly recalibrate campaigns to focus them on specific market personas and opportunities, that is also the problem. There is so much data, such an enormous volume of one-off (siloed) consumer opinion and behavior coming at us so fast, that accurately interpreting intent and sentiment now requires advanced analytics. Marketers, advertising executives, and product developers can't drink fast enough from the fire hose to anticipate each new direction and trend.

Unstructured Data and Advanced Analytics

Today, it is *unstructured* data that is seen as the key to interpreting and understanding the online sentiment and intention within the massive numbers of posts, tweets, pins, and videos from countless online sources and channels. Unstructured data—text, audio, imagery, video, and posts and links from Web sources—and the integration of this information is the path to ultimately developing brand and other business strategies for improved customer experience, brand equity management, higher revenues, and mitigated business risk. Now comes the loaded question: among all the noise, how do we read it, catalog it, and ultimately understand it?

Ultimately, the goal of advanced analytics is to empower business users to potentially influence strategic imperatives by using a contextualized alerting system. What does this look like? Imagine a human resources manager at a midsize technology organization tasked with understanding high turnover rates within a specific software development group. With access to an advanced analytics solution (preferably one with an elegant, made-for-the-masses user interface), the manager could construct a data module that scours both external and internal discussions about the organization, identifies sentiment on a variety of trending topics, and observes certain clusters of conversation types. The HR professional, discovering employees in a certain age bracket are concerned about long-term job stability or rumors about diminishing retirement benefits, could act on these combined internal

and external insights and feedback to quell fears, make valued employees feel more secure in their positions, and create better corporate alignment.

The list of similar uses of advanced analytics is never-ending, with tactical practitioners able to focus on a myriad of listening programs. A marketing intern could track a new product release and provide upper management with campaign insights for possible adjustments. Product developers could see quality issues in real time and revamp accordingly. IT could consolidate social tools and improve processes by taking a more holistic view of the customer.

Organizational needs are moving beyond traditional descriptive and predictive analytics to advanced analytics, which rolls predictive modeling, clustering, affinity analysis, and optimization into a more robust, prescriptive analysis formula.

Even with the right technology, it is a big job. The sheer volume of available material, combined with the pace of creation, demands the ability to shift constantly in real time, all with a human-language understanding of the content and context. Organizations want to move beyond traditional business intelligence reporting, descriptive analytics, and diagnostic analytics to advanced analytics practices such as predictive modeling, clustering, affinity analysis, and optimization. In addition, analytics use has expanded beyond analysts and data scientists to reach ordinary users.

To achieve this delicate balance, keep the following requirements in mind:

- **Real real-time results:** Staying up-to-date with the pace of social media is taxing but it is key. These days, five minutes ago is often too late, especially if your competition is monitoring the same data.
- **Filter spam:** A key issue with analytics for business users is the vast amount of spam that comes in many analytics reports. Without effective spam filters, the results are often watered down or even—sometimes drastically—skewed.
- **Set alerts on a topic's volume, sentiment shifts, trends, and significant influencers:** The days of having a dedicated employee to monitor analytics streams 24x7 are gone, and being able to stay abreast of important developments *as they occur* with an alert is important in a business world where people often wear many hats.
- **Track and quote metrics for volatility, sentiment, mentions, followers, and trend scores:** Users must be able to track and view results from various angles to dig deeper into specifics.
- **Configure on-the-fly, business-themed tagging:** Tagging is still relevant and topics shift quickly, so being able to easily edit your tags and searches without interfering with algorithms is necessary. For example, the most popular search suggestions to appear as you type "iPhone" may include "iPhone 5S," "iPhone 6," and, in time, "iPhone7," etc., making Boolean search syntax a thing of the past.
- **Avoid scrimping on the user interface (UI) and visualizations:** The back end of analytics is complex, but end users don't have to be reminded of that every time they look at the screen. Effective UIs allow for several views (volatility, sentiment, etc.) so users can easily view and track real-time results in an attractive interface.

Finding the Right Mix of "Too Much" and "Not Enough" to Provide the Most Accurate Results

With unstructured data, enterprises have a new window into online sentiment. Where it was once the job of

an intern to compose perfectly parsed Boolean queries in a vain attempt to gather intelligence, we now see artificial intelligence, machine learning, and sentiment analysis—the backbone of natural language processing (NLP) technologies—used to deliver real-time, contextual, and intent-aware social analytics.

We must move to the next level of analytics: text analytic tools that understand context.

However, NLP technologies raise a new issue. With so much data crunching power at our fingertips, results are often over-simplified. Text analytics tools often return "black or white" or "yes or no" results. What users desire is a technology that goes beyond the basic levels of text analytics to also measure deep sentiment and intent. For example, there are big differences between Apple the electronics company and apple the fruit, Sprint the company and sprint the verb, McDonald's the restaurant and McDonald's the farm. It seems obvious enough, but many users are not aware of this problem or equipped to handle it (unless you delve into heavy Boolean query styles).

To solve this problem, we must move to the next level of analytics: text analytic tools that understand context. We have an array of ideals to consider, including understanding the profile of the author (gender, age, demographics, etc.) as well as gauging the social channel. After all, LinkedIn is a vastly different conversation forum from Facebook, where one statement might be read with serious professionalism and the other with sarcasm.

Take, for example, slang. If we say, "Beyonce's shoes are SICK," or "He is MAD for BrandX Ale," or "I got WRECKED at the pub last night," many filtering and analytics tools won't pick up the positive tone implied—but rather score the dialogue as negative or spam.

Another threat to accuracy is using analytic tools that only define people as influencers based on a score—such as Klout or Twitter do—rather than delving into the context of who the influencer is and what they actually *feel* about a specific domain.

Sentiment accuracy is a big sticking point because users need to filter out spam and other irrelevant results to truly measure campaign benchmarks, trends, and effectiveness more accurately. They must be able to surface intelligence from the noise, and to accomplish that, natural language processing is required.

Natural Language Processing's Role in Sentiment Analysis

Natural language processing is a combination of machine learning, artificial intelligence, and semantics that helps computers derive meaning from free-form human communications online or via machine.

However, teaching computers to accurately understand how humans speak and write is just one piece of the NLP challenge. Equal consideration must be paid to extracting the context embedded with online conversations, including who, what, where, when, and why. Context provides a deeper picture of content and helps brands reach their ideal customers by moving beyond using only high-level social media metrics to define user profiles. This is the gold standard in discovering new opportunities and threats, and although such metrics are immediately available across most social channels, their sheer volume makes it a challenge to organize them with accuracy and value.

Some analytics tools that can't grasp and assign context default to using a "sentiment lexicon." This is a system of assigning positive or negative scores to specific words, and then assigning a high-level, generalized meaning to an entire post, tweet, review, or article. This approach has obvious consequences, most of which ignore the need for context to accurately define a sentiment as positive, neutral, or negative.

For instance, some words in our vocabulary are neither positive nor negative, and analysts may incorrectly assign

Question	Do you think I should close my account?	Should I buy an iOS or Android phone next?
Negation	I'm not going to close my account.	I don't like the new S6.
Conditional	If someone doesn't call me back, I'm closing my account	If my iPhone breaks again, I'm going with Android
Intent	I'm closing my account tomorrow	I'm heading to the Apple store tomorrow
Past	I closed my account yesterday	I threw away my iPhone yesterday
Urgent	I'm closing my account ASAP	My iPhone 4 just died and I'm upgrading tomorrow
Indefinite	I might close my account—I'm not sure	I'm on the fence about a new iPhone 5S
Augment	I *have* to close that account	I am *so done* with Apple
Request	Please close my account	I can't wait to upgrade to the new iPhone 6
Diminish	My account is not meeting my standards	My iPhone just isn't doing it for me
Suggestion	I'd close that account	Have you thought about Android?
Recur	I had to close the account because they kept billing me incorrectly	I ditched my iPhone because it kept crashing
Command	Close my account!	Fix my new iPhone!

Table 1: Natural language processors must account for tone of voice when interpreting content.

different degrees of sentiment to specific words. Take, for example, this sentence: *I was looking forward to eating at the restaurant, and even though all the ingredients were first grade and everything on my plate looked beautiful, it didn't taste like anything special*. Most words would be assigned a positive score, yet the overall sentiment is negative, making ambiguity yet another hurdle that analytics tools need to overcome.

This is why NLP—which allows computers to understand, process, and analyze freeform text as well as the subtleties of words—is critical to creating contextual understanding. In fact, the interpretation and categorization of content helps us understand the context that customers use and to ultimately know the nuances of language in order to reliably establish the *context* of what is being said: When is "apple" a fruit and when is it a technology company?

Natural language processing ensures that tone of voice affects the interpretation and categorization of content to consistently and reliably establish whether it is a negative, positive, neutral, or mixed sentiment—as opposed to merely +/-, positive/negative, or yes/no responses. When employing NLP, consider the variety of sentiment "tells" shown in Table 1.

What's Next: Distributed Analysis and the 360-Degree Customer View

Data analytics—specifically sentiment analysis combined with natural language processing for a true and contextual understanding of online declarations and conversations to create the "beyond sentiment" piece that we discussed earlier—will ultimately lead businesses to success.

When this occurs, it will be because data research has evolved into "distributed analytics" designed to scale and accommodate explorations of massive data volumes to ultimately organize it in a meaningful way. The end goals—profiling and prediction—will rely on processing these massive amounts of siloed, distributed data with the understanding that each individual piece of data, or set, must also be viewed as a communication and interaction with other bits of data.

Let's revisit our profiling idea. Although everyone competes for accuracy in degrees of sentiment, its delivery can only be found in a 360-degree view of the customer. For example, consider insurance and claims analysis. Actuaries depend on profiling, taking into account major and minor life events such as marriage, birth of a child, home purchase, pet adoption, change in health, a new job, or pending retirement. Each event links together to build a profile of a prospective customer, and every

customer is linked to (and influences) other prospective customers in a set, thereby revealing the deep human relationships within collective data.

Another notion to consider is time. History and the past further allow us to create a complete story and develop an outline of when a customer's history began, when his relationship with the company will peak, and when it is likely to end—and why. What we are seeking is a story with a beginning, middle, and end. Discover this and you are more likely to predict the end and base business decisions on that timeline. Such predictions can alert a business to what is coming—whether good or bad. What happens when these contributing factors get wrapped into an analysis path?

Data analytics is entering a new stage—one that will string together historic data with current events, nuances, and (until now) largely unstructured pieces of data to create The Big Picture.

Consider a mother or father researching kindergarten options for a child. A given scenario will hold a huge data set that will drive the conclusion or selection, involving general demographics, personal and professional data, even political affiliations, religion, and future objectives and goals. What school will be chosen, and what factors influence the selection? It could be any number of data sets, such as neighbors with children of the same age entering the school, a new job that requires flexible pick-up hours, or even the online search for a new car to replace an aging, less reliable one.

Much of this information is going to be siloed, but the conclusion is a collaborative result of each individual data set combined and bounced off the others. The same can be said for social conversations that may appear to take place in a vacuum, but in reality are transactional decisions based on a never-ending flow of influencing factors. Winners in the social listening space will tie all of this data together across business units.

The same holds true for analytics—sentiment specifically—that demands increasing integration of the sorts of unstructured data that can only be found in a soup of patterns, including history and demographics. Data analytics is entering a new stage—one that will string together historic data with current events, nuances, and (until now) largely unstructured pieces of data to create The Big Picture.

The result is a trend prediction created by distilling information from mass quantities of individual interactions.

The Holy Grail

Increasingly, the integration of unstructured data, sentiment analysis, and NLP needs to play a role in painting an accurate picture of online conversations and data to help with a top priority for customers: discovery. In other words, "help me discover what I don't know."

Although the social media world is currently (still) basing most of its data analysis on follower numbers (such as Klout scores and LinkedIn profile networks), some technologies are poised to exploit more accurate methods of social analytics and delve into dynamic search, context space influencing, and more.

These technologies will cause a mind shift in the market and lead the new practice of monitoring conversations as they are meant to be interpreted based on who is speaking, who they are speaking to, what their collective and individual histories and backgrounds reflect, and all other manners of unstructured—but available—data. When this happens, businesses will finally be able to distill context and evaluate sentiment to determine influence and ultimately seize the Holy Grail: understanding intent. ■