Smarter data

The pace of data collection can outgrow that of useful, extractable information, so for information to provide insight it must be interpretable, relevant and novel

By Dr Michael Wu, *Lithium Technologies*

Data is only as valuable as the information and insight we can extract from it. It's information – not data – that helps us make better decisions, and it's insight that brings competitive advantage. Logic would have it that because information and insight are derived from data, the more data we collect (the bigger our data gets) the more information and insight we'll be able to glean. Not so.

BIG DATA MYTH #1: DATA AND INFORMATION ARE THE SAME

When you back up your hard drive, your data volume doubles; but will you get twice the information? The answer is pretty obvious that you actually won't gain any extra information from backing up all the data in your hard drive, even though you've doubled your data volume.

The problem with data is that as it gets bigger, it doesn't necessarily yield proportionately more information. In fact, the opposite is most often the case. The more data we collect, the value it brings as a proportion of all the data we house will typically go down. As we hoard more tiny bits of data, the percentage of it that matters dwindles.

This is partly because much of the data we collect is statistically redundant. Note that statistical redundancy is not the same as data duplication. Duplication is only the most trivial kind of statistical redundancy. Statistical redundancy between two data sets means that there exist some common facts we can learn from both data sets. For example, consider the following two small data sets – two tweets:

- **User A:** Apple revolutionises our mobile communication experience.
- **User B:** Steve Jobs led one of the most innovative companies.

Even though these two tweets don't even have a single word in common, there is some statistically redundancy, because we can learn that Apple is a very innovative company from both tweets. We extracted this same bit of information from both tweets with our brain by interpreting what we read. Because there is some information common to both tweets, the two tweets have some statistical redundancy, albeit only a small amount.

This kind of statistical redundancy is common in all data, but it's especially prevalent in social media. With all the sharing and retweeting going on across the social web, it's easy to see how the data we collect can get very big indeed while the extractable information grows much smaller.

BIG DATA MYTH #2: ALL INFORMATION PROVIDES INSIGHTS

Although the amount of information may only be a tiny fraction of the big data we capture, the amount of insight we can glean from the data is even more elusive. This is because not all information is insightful. The two tweets above are not really insightful because the information we can learn from them doesn't tell us anything that we don't already know. Information must satisfy the following three conditions in order to provide insight:

- **Interpretable** With so many channels, media and data types to interact with, much of our data comes to us totally unstructured. In fact, much of the data we see isn't even interpretable. Consider this sequence of numbers: 432, 143, 84, 9 and 156. They could be your Twitter followers, the number of likes your last Facebook post garnered, or the number of community posts made last week. Without metadata – data that tells us what this data is – there is no way to interpret what these numbers mean. Only interpretable data can offer insight.
- **Relevant** Information must be relevant to be useful and valuable. Especially when it comes to social media, the signal-to-noise ratio – the proportion of relevant data to irrelevant data – is often quite tiny, making most of the data we collect seem unnecessary. But it's important to remember that relevance is subjective and contextual. One man's signal may well be another man's noise. And what might look like noise today could go on to become signal tomorrow. If you plan to travel to New York City next week, then the traffic and weather in New York City will suddenly become very relevant. Once you return home to San Francisco, that same information will become irrelevant. The upside of data is that it has a long shelf life and it's very versatile. Since we can never know what might turn out to be relevant in the future, it pays to hold on to our data.
- **Novel** Information must be novel to be insightful. It must provide some net new knowledge – knowledge that we don't already know. Clearly, this criterion is also subjective. What is known to one party may be unknown to another and vice versa. This uneven distribution of novel information is how insights offer competitive advantages to businesses. Note that the novelty requirement doesn't automatically guarantee the discovery of insight, because it's only one of the three necessary conditions. However, the absence of novelty does preclude the possibility of finding any insight in the extracted information. A peculiar property of the novelty requirement is that information can only be novel once. Once an insight is found, it's no longer new the next time you find it again. As a consequence of our ever-growing understanding, as we accumulate more knowledge and insight from Big Data, insights (i.e. novel information) become harder to discover. Therefore, the insight we can glean from Big Data will always be a tiny and continually shrinking subset of the relevant information.

Clearly, data does not equal information, nor does information equal insight. In fact, the volume of insight gleaned will always be a tiny fraction of the information it comes from, and the amount of information available will again be a small percentage of the sheer volume of data collected. That said, as a single grain of rice can tip the balance, one bit (1/8 of a byte) of insightful information may be the difference between victory and defeat. So, even though insights are extremely rare in Big Data, you don't really need many insights to have a significant impact on your business. When it comes to data, business must retrain its mindset away from 'big' – simply collecting as much as possible – and towards 'smart' – doing the analytics to extract information and insights from the data.

So what types of analytics are needed to extract information and insight from Big Data? There are three categories:

- **Descriptive analytics** – summarise past data we've collected (note: all data, even real-time data, are past data that tells us about events that have already happened).
- **Predictive analytics** – use data that we've collected to infer data that we either can't collect, didn't collect or haven't collect yet.
- **Prescriptive analytics** – using past data on actions and outcomes as feedback to prescribe actions that would best achieve the desired outcome.

These analytics are all useful for the extraction of information and insight from big data and no one is more important than another. However, it does take some analytics maturity and data savviness to move from descriptive towards predictive and prescriptive analytics.

Currently, more than 80% of business analytics are descriptive. When it comes to social analytics, the percentage is even higher – approximately 90%. Most of the remaining 10% are simple predictive analytics based on linear trend analysis. There are only a few sophisticated predictive analytics available today that try to infer consumer sentiment, intent, influence, etc. from their social media interactivity data. Since social analytics is still a young discipline, there are virtually no prescriptive analytics available today.

Regardless of what types of analytics we perform, it is important to visualise the results in a way that is intuitive and actionable. If the result of the analytics we perform doesn't help the business make better decisions, then all our analytic computation is spent in vain. If decision-makers can't easily take actions from the results of our analytic computations, then we've failed to realise the value of Big Data.

> "Regardless of what analytics we perform, it is important to visualise the results in a way that is intuitive and actionable"

While Big Data technology remains relatively cheap, we can expect the Big Data drum to continue beating. Although Big Data cannot guarantee the revelation of huge numbers of valuable insights, increasing the volume of data we own and can access does increase the odds of finding those insights. However, it's important to remember that while the technology isn't expensive, the total cost of ownership involves more than just the technology. Much time and talent is needed to carry out the analytic process of information extraction and insight discovery from Big Data.

While Big Data technologies provide the infrastructure for capturing, storing and processing unprecedented amounts of data, they are only enablers. What businesses really need is information that helps them make better decisions, and insight that gives them the competitive edge. Presently, these are only attained through human endeavours due to the creative nature of analytics and data visualisation.

This article was first published in Admap magazine February 2014 ©Warc
www.warc.com/admap