Print Article Close Window

From: www.cio.com

Social Data Warehousing Is Worth the Trouble

– David Taber, CIO

April 18, 2013

Classic data warehousing collected enormous amounts of relational data from sources across the enterprise and then correlated it to create more meaning than could be seen in any one system. Most of the data was purely relational, and most of the inferences were pretty straightforward, even if the joins were tricky.

But when you're looking at social marketing, sales 2.0 and social CRM, you need to pay a lot more attention to time-series data and the interactions across social networks. These can all mean a ton of data.

First, with behavioral scoring, a marketing automation system not only needs to track every email you've sent, but also every response—including all the pages a user visits, all the cookies that have been dropped, every phone call and the click path that led to a purchase. The system needs to track almost the same amount of data for anonymous visitors as it does for leads. Even small companies may be recording millions of data points a month.

Second, with social networking, it's not enough to know which social networks someone belongs to. The high ground is creating graphs of the social network based on patterns of emails, phone records and social postings to help you understand who are the mavens or connectors who have the most influence on the community.

Analysis: How CIOs Are Getting New Value Out of Social Media

These graphs also help you understand the most direct or credible way to connect to or influence a prospect. While the simple connection matrix isn't very large, the social-influence matrix may become N-dimensional and prodigious.

Third, instant messaging and other social feeds can be helpful to track audience sentiment and support lexical analysis. But these are the height of unstructured data, particularly if you include attached files. However, these can be important to record if you are interested in analyzing brand mentions or logo appearances in stills and videos.

Social Data Challenges Reflect Quantity *and* Quality

It's not just that each of the feeds described above is big. It's the need to maintain time sequencing and correlate events across several media. That leads to dreaded combinatorial explosions.

Case Study: Social Media Marries Big Data at Wedding Retailer

The obvious answer is to do most of your analysis on extracts or tallies, rather than on the underlying record-level details. That works as long as you're doing fairly stable analyses, where you can

pre-determine most of the queries and all of the extracts. Soon enough, though, somebody will have a follow-up question that requires examining the detailed data, so you'll need to have tools that can drill down below the extracted summaries.

The economics of the cloud, and the speed of deployment there, can make for compelling advantages. There are now a number of solid BI tools available only in the cloud, and users of cloud-based operational systems are increasingly using SaaS for their data warehouses. Cool.

With social data, though, it's only the extracts that can be practically handled in a pure cloud warehouse.

The underlying details—for ad hoc queries, hypothesis testing, and extract formulation —will almost certainly have to be done with on-premises databases. Fortunately, disk and memory capacities continue to fall while capacity expands. (The laptop I'm writing this article on, for example, has more than a TB of internal disk space).

How-to: 5 Tips to Find and Hire Data Scientists

The real cost of the on-premises warehouse, though, will be the software and the data analyst. While there is good news in terms of analytical power, neither the people nor the software is likely to come down in price any time soon.

Prune Social Data Early, Often

I'm usually a card-carrying data pack-rat, but with social data warehousing, there's not much point in keeping detailed data around for too long.

The first reason is information value. Much of your social data becomes obsolete as the rules of the game continue to change.

- Cyber-social mores are evolving rapidly. There's not much profit to be gained by understanding how people interacted in MySpace or SecondLife. Let's face it: Some social-network behaviors are vulnerable to fads. Leave the pure research projects to the academics.
- Advertising platforms and tactics are evolving rapidly, particularly for mobile audiences. The threshold values for click-through ratio—and the importance of it in understanding conversion ratios—are hardly fixed.
- Your competitors' actions affect your results, and your goals will change from year to year. Given the size and complexity of the data space, it will be almost impossible to normalize analytics over the long term. There's not much hope of discovering universal coefficients and algorithms that will be good over long periods, so focus on the here and now.

Analysis: 4 Barriers Stand Between You and Big Data Insight

The second reason is signal-to-noise ratio and the costs of processing it.

- A significant proportion of the social data you collect will be noise. The initial data points may look promising, but in many cases the user you're tracking took no action or simply disappeared from view. In some example data sets, we were able to throw out all the data from more than 95 percent of the prospects we were tracking.
- Even if you get your data warehouse software or service for free, there's a non-zero cost for the time and effort of managing—let alone analyzing—the avalanche of data. We've seen some real-time systems that were not able to even delete more than a month's worth of data per query. Imagine how long it would take to aggregate all your data.

Social data warehousing is so new that we have to reinvent the practice, as well as the tools, to be effective. Make sure your social data warehousing project has a clear (and probably short-term) goal, as well as tight management, so it doesn't become a money pit.

David Taber is the author of the new Prentice Hall book, "Salesforce.com Secrets of Success" and is the CEO of SalesLogistix, a certified Salesforce.com consultancy focused on business process improvement through use of CRM systems. SalesLogistix clients are in North America, Europe, Israel and India. Taber has more than 25 years of experience in high tech, including 10 years at the VP level or above.

Follow everything from CIO.com on Twitter @CIOonline, Facebook, Google + and LinkedIn.