WSJ.com THE WALL STREET JOURNAL. ONLINE

A Good Sort

Jeremy Wagstaff, Wall Street Journal Online

February 24, 2006

http://online.wsj.com/article/0,,loose_wire,00.html

Here's this week's tip. If you want to sort the wheat from the chaff -- whether it's separating email from spam, hot stocks from duds, or great movies from bombs, you'll need the help of an 18th-century vicar.

Take, for example, the experience of Matthew Prince, chief executive of Utah-based antispam consultancy Unspam Technologies Inc. Hooked on the annual Sundance Film Festival since 1996, he has had the same problem facing everyone who attends any big cinema festival: Which of the 200 or so films being shown are worth watching? So he and a group of friends began trying to find ways of picking the best ones based on reviews of the films being screened. One day, chatting with a fellow software engineer, they both realized that picking the best films was really the same problem as deciding whether an email message was spam.

What's with that, I hear you say? Let me take a moment to explain why most of the world's spam -- and there's a lot of it -- doesn't end up in your inbox. It's all because of Thomas Bayes, an 18th-century English vicar, who came up with a theorem to calculate the probability of a future event based on past events. His theorum forms the basis of modern-day spam filters used by most Internet Service Providers and email services. Put simply, if a piece of spam you receive contains the word "Viagra," chances are high that subsequent emails you receive containing that word also will be junk email. A Bayesian filter will inspect all the words in an email -- including hidden formatting, the headers and other telltale signs of spam, and assign a probability of the email message being junk. All you have to do is to train the filter by showing it a handful of junk and email messages, telling it "this is one is spam, this one isn't" and it starts quickly filtering out the rubbish.

Mr. Prince's bright idea was to apply the same filtering technique to film reviews. Would it be possible, he wondered, to throw film reviews of the past few Sundance festivals through a Bayesian filter and see whether it could pick the likely winners? The U.S. Sundance festival, which is a leading showcase for independent cinema, releases a guide to the films being screened every year. Mr. Prince and some colleagues gathered 10 years of guides to more than 360 Sundance films. Based on the individual film's success at the festival and subsequently, each was assigned to one of three categories, or baskets: Below average, average and above average. Their findings gave birth to the Web site (http://deconstructingsundance.com).

What Mr. Prince and his colleagues found was that, among other things, words were a pretty good indicator of success. But not necessarily the words you might expect in a review: Best. Fascinating. Emotional. Inspired. Great. All are, in the words of the Deconstructing Sundance Web site, "the kiss of death" for a movie. Riveting, for example, appeared in 46% of reviews for what turned out to be below-average movies, as opposed to 22% of above-average movies. How so? Why would a reviewer call a dud "riveting?" Mr. Prince has his own theory: "Maybe writers, when they struggle with something good to say about something, revert to adjectives like 'riveting' rather than actually describing the movie in a more tangible way?"

Pretty neat. But why stop there? If an 18th-century cleric can help you figure out which movies are going to make it, why not use the technique to predict other things, such as stock market movements, blood clots, or volcanic eruptions? Well, actually, there are people thinking like this. U.S. shopping search engine Shopzilla.com uses a Bayesian filter to sift customer emails according to topic and, where, relevant, fire back canned responses.

But what can this do for you? Well, if your ISP or office network isn't filtering out your spam, you can set up your own Bayesian filter. I suggest going with POPFile (http://popfile.sourceforge.net) a free, all-platform version of a commercial product called PolyMail developed by John Graham-Cumming. (It was he and POPFile who made the whole Deconstructing Sundance thing possible.) It's relatively easy to set up.

I've used POPFile for a few years and it's kept the spam at bay. Recently I decided to make it work harder. As with Mr. Prince and his crew, I felt that if the software did such a good job with spam, why not let it sort all my email out for me? Email's big problem, you see, isn't just about filtering out spam. It's about sorting everything that comes in, so it doesn't all land (and usually stay) in one big oversize inbox.

My advice is to set up two baskets -- say, Personal, and Work -- and a Bayesian filter will quickly figure out where your email will go. Instead of having to write a rule for every sender, or for every email with the words "Loan Shark" in the subject field, you can just teach it where a few sample emails go, and then leave it alone. I'm now experimenting with three baskets: 1) what I need to deal with now, 2) stuff I can save for later, and 3) stuff I'll never need. So far it's working pretty well.

Of course, in all honesty, we don't know quite why the Bayesian system works. It just does. Expect the good vicar's theorem to spread beyond spam control to other applications on the Internet.

Oh, and the Deconstructing Sundance project got it right in shortlisting some of the potential winners at this year's Sundance festival, which finished a few weeks ago. They tracked the buzz on two films, for example, that ultimately won the festival's two top awards: "Quinceañera" (dramatic) and "God Grew Tired of Us" (documentary).