"We are what we search for," says Andreas Weigend. He says that a powerful compression of people's lives is encoded in the list of their search queries.

I SEARCH, THEREFORE I AM

Jochen Wegner

As the former chief scientist of Amazon.com and one of the world's leading data mining experts, Andreas Weigend and his opinions are widely respected. If you missed his talks at SAS Forum International or M2004, the world's largest data mining conference, you won't want to miss him now. In the following article, Weigend explains how data can be used to accurately predict customer behavior. The question-and-answer piece is reprinted with permission from the Oct. 4, 2004, issue of *Focus*, the most widely read German-language news magazine, that boasts more than 6 million readers.

Are all people the same?

«Weigend» No, of course not. People are individuals, most highly idiosyncratic. And this is reflected by the multitude of traces they leave in our connected world. However, it is surprising how accurately their behavior can often be predicted. This is a fundamental paradox in my discipline.

Until earlier this year, you were the chief scientist of Amazon.com. Amazon.com is known for individual product recommendations that are often right on target. You are now advising a number of companies that have rich user data, including match.com and gay.com, two leading online dating sites. What is it that makes people so predictable, even in such difficult areas as literature taste or partner selection?

«Weigend» We all believe we're individuals who make deliberate decisions, whether it's about buying a book or finding a partner. However, most people are way more predictable than they believe. If they are in a certain situation, they will react in a certain way. If you follow customers over time, you discover strong regularities, for example, in their information-foraging behavior. Additionally, short-term human behavior often has indicators that make it much more predictable than long-term behavior.

Could you give an example?

«Weigend» If in the past year you only purchased biographies of Russian czars, then, indeed, you might do just the same on your next visit. A bookseller can use this information to alert you if a new biography in this area has been published. In many instances, however, your current behavior – say, what you are looking at right now – is an additional, highly relevant source of information. In this case, recommendations are based on the purchasing behavior of all customers – your past is irrelevant. Amazon.com manages a giant matrix relating millions of products to each other. An entry, or a cell, in this matrix essentially counts the number of individuals who purchased the two specific products that correspond to this cell. This is the basis for generating recommendations in response to your current click. By the way, this type of recommendation works just as well if you are visiting the site anonymously since your past behavior is not used by the algorithm.

Can you really predict my next action by simply evaluating my current behavior?

«Weigend» Yes, quite well. But more importantly, not only predict, but I can use this information to spur action. I spent the 1990s consulting to major Wall Street firms. There, most of the work is done as soon as you've built a model that predicts reliably. Why? The corresponding action is simple: Buy if the price goes up; sell if it goes down. (This of course ignores risk and portfolio aspects.) In e-business, making good predictions is only the very first step – which of the millions of items will you show, where on the page, at what price, etc. Determining the best action that follows your prediction is where the hard work lies. Why? In retail, you control the response of the system. In finance, you have no way to influence the system, unless you are willing to lose a lot of money.

Back to online shopping, give me an example of how the item you show depends on what you predicted.

«Weigend» Consider the modalities of a user. Is he in a hurry or just killing time? There are many other attributes. We are multimodal. We are interested in many things. Let's say you'd like to buy a present for your mother. In this case, your interaction history with the site is of far less consequence than the fact that you are clicking on a specific pair of warm winter socks. This piece of information is then used to predict what you will do next, and in this game, the site will act upon it by showing you things that might fit your current mode. This assumption results from reading out the matrix I mentioned before. The precision of aggregated consumer data is extremely high because millions of customers refine this data on a daily basis through their purchasing behavior.

"Data can be used to accurately predict customer behavior."

How would you know whether I want to buy socks for my mother or a turkey baster for my Aunt Tillie even before I click anywhere?

«Weigend» Well, the question is, what data do we have? It's often more important to creatively invent new data sources than to implement the latest academic variations on an algorithm. A Web site often knows which site you visited before. For example, when you click on a link that leads you to Amazon.com from a site on knitting, Amazon.com knows that. And you are likely to be in a different state of mind if you came from a search engine through a keyword, or from a shopping comparison engine, or if you typed "www.amazon.com" in your browser. It also matters whether you keep clicking and clicking and clicking, or whether you enter something into a search box. And if you conduct a search, it depends on how many results you get back. Looking at the search terms that people use and refine can be a revealing view into the mind and the soul of a person. A powerful compression of our lives is encoded in the list of our search queries. We are always looking for things that we don't know, but would like to know. We are what we search for. *Quaero ergo sum* – I search, therefore I am.

This is a deep point, thank you. But, more practically, how can a company use an analysis of searches to increase sales?

«Weigend» It's easy to determine how specific a search is. If only, say, two results are returned, the customer seems to know pretty well what they want. However, when thousands of results are returned, then they need to be guided. This is an example of how a simple number – how many results came back in response to the query – indicates where the customer is in the process of creating and maintaining product space awareness.

So you're saying you can figure out what people want even before they know it themselves?

«Weigend» Stated preferences differ from revealed preferences for a number of reasons. My approach is to measure both and then model their relationship. At Amazon.com, I conducted an online survey asking the customer what they planned to do in their current visit. We then compared their responses with their clickstream behavior. On the one hand, only one-third of those who stated they wanted to buy something actually made a purchase in this session. On the other hand, half of those who made a purchase came without the explicit intention to buy – what an opportunity for the marketer! Such numbers are of course only a first step – it was fun to click through people's sessions and hypothesize why people who wanted to buy didn't, and the other three combinations. I am a physicist by training and love doing experiments, and one of the great things about the Web is that it is a huge laboratory that cuts down time scales by orders of magnitude.

What do you mean by time scales?

«Weigend» I mean the time it takes, roughly, to establish whether a cause has an effect or not. A hundred years ago, in agriculture, scientists needed to wait for a year until they saw how a new set of crops were doing. And even now, in the physical world, it takes months to figure out how a new flavor of toothpaste is doing. On the Web, if you change a property of the Web site, such as the placement of an offer, you know the effect within minutes. Another example of action on fast time scales is our ability to characterize the situation a customer is in at a given moment. A company can learn about an individual's state of mind by presenting offers that distinguish between different states and measuring the customer's response.

> **Looking at search terms people use and refine can be a revealing view into the mind and the soul of a person.** – Andreas Weigend

Can you give an example?

«Weigend» Amazon.com had a feature called "Gold Box" that presented 10 items sequentially, one after the other. After each item, the customer had the choice between buying that item or not buying that item. If he did not buy the item, a different, new item was presented. Terry Odean at Berkeley, Itamar Simonson at Stanford, and Dan Ariely at MIT suggested a new design, based on their rich intuitions. There might be a computer mouse and an outdoor grill. The customer can then keep the item they prefer. The other item is then replaced by a new item. In the given example, if someone prefers the grill, you might infer that they are more interested in another house or kitchen item than in another computer item. Interestingly, even when items are presented in a random sequence, Wendy Liu and Itamar Simonson found that people are 15 percent more likely to buy if they are given the choice to hold onto items for a while. And then data mining and machine-learning researchers enter the game and cook up algorithms for selecting that specific item to present.

That's plausible, but a bit vague.

«Weigend» Sure, let me be more precise. When I offer you in a second step again the grill and offer as an alternative a pair of hiking boots...

... then you also know whether Aunt Tillie wants to grill or hike the Appalachian Trail.

«Weigend» Exactly, and you can play that game 10 times. The number of rounds that you stay in the game will say something about your current time sensitivity, your curiosity, and perhaps about your intention to buy.

But aren't most customers unlikely to play the game for 10 rounds? Are there other ideas?

«Weigend» Actually, many customers do play the game for its entire length of the 10 rounds. And yes, there are new ideas: A promising new concept is that of the network value of a customer. It's one thing how many dollars you've spent buying stuff for yourself at a certain company. It's a totally different thing how many people you've influenced to buy things from that company. How influential are you? Measuring the network value of a customer can be worth big bucks.

And how do you measure it?

«Weigend» You might have seen that Amazon.com lets you recommend to your friends a book that you just bought. If at least one of your friends buys a copy of the same book, you both get a discount. Some people have a high success rate of getting their friends to buy, others don't. This feature can thus be used to characterize influencers. Other inputs, including reviews, can be used to compute the amount one customer influences others. This is called the network value of a customer.

Isn't this a privacy advocate's worst nightmare?

«Weigend» Yes, when looking at it superficially. However, for me, the ultimate goal is to educate people. I want them to understand that there are trade-offs, and learn what these consist of. I want them to make their decisions knowingly and consciously. For example, if you buy an electric toothbrush from a retailer and explicitly tell that retailer to destroy all purchasing information, if the toothbrush then breaks and you try to return it, what do you think the retailer's response will be? I want people to be aware of the spectrum of options available to them, and the consequences of choosing specific ones. I want to empower people to manage their data responsibly, including decisions on their privacy. A company must honor these decisions and treat its customers with integrity and respect. I hate it when a company tries to manipulate its customers, pulling the wool over their eyes.

Imagine a data miner knows the location of my mobile phone at any given time. And that data miner knows that I read an article or clicked on an item. Add to that the fact that the miner knows what street I happen to be strolling along. Might I not also be interested in a special offer from the shop just coming up? Should that shop send a message to my phone?

«Weigend» Why not? But I want to be in control of whether or not I want to receive that message! And please don't be stuck in the pathetically simplistic binary mode of my phone being on or off. I want my phone to understand my preferences, negotiate with the store and decide whether or not to deliver the message. Technology can facilitate a simple economics exchange: There are two sides to the potential deal of the store sending me a message. The store

and myself, in the current situation. And the simple question is whether the amount the store is willing to pay for my attention, such as the number of dollars deposited into my account if I allow delivery, is below or above the current threshold I have for a store like that.

How does your phone know that?

«Weigend» By modeling my past behavior from the communication data it has collected. It is a rich source of data about me, because it is always with me. In some sense, it knows more about me than any human. It always knows where I am, via GPS or the location of the nearest base station. It can hear – all mobiles have microphones. It can see, using the built-in camera. It knows who I know, since I have my address book on it, and remembers the times of interactions, through voice, e-mail, chat, messages. It knows my calendar, and thus can figure out whether I am late already for my next appointment, or potentially have time to check out that store. And if I searched for the store earlier today, it might even be willing to pay the store for helping me find it more quickly, rather than collect a charge for a bit of my attention.

Is this the future?

«Weigend» When we started, I mentioned that human behavior is more predictable than we think. That statement was based on the pretty narrow part of our lives, the clicks and purchase data from an online store. I then discussed other ways we express our preferences: the searches we run, the paths we take through the Web, and gave you my little phone example. Communication is what enables it all. But this is only the first step. Wal-Mart is testing at its distribution centers its deliveries tagged with radio frequency identifiers (RFIDs). It is expected to save Wal-Mart more per year than Amazon.com's entire revenues. These are little chips, the size of a grain of salt, costing a few cents, that uniquely identify an item. Like the sweater you are wearing. Or the toothpaste in your bag. In a few years, we will probably remember those book recommendations you asked about as a pretty minor thing of the early days of data collection and mining. The key is to educate people to understand the inherent trade-offs, to empower them to use technology in the way they choose, and to fully respect their decisions. ■

What makes Weigend tick?

As the chief scientist of Amazon.com, Weigend developed quantitative approaches for situation-based marketing, generating customer attributes, and computing customer network value, the amount a customer influences others. As educator, he currently teaches the graduate course, Data Mining and Electronic Business, at Stanford University. He is a sought-after speaker at international conferences and in executive education. As author, he published more than 100 scientific papers and six books in the areas of particle physics, computational finance and machine learning. Until 1999, he was a full-time faculty member at New York University's Stern School of Business. As consultant, he works with data-intensive organizations creating strategy based on measurement and behavioral analytics. Clients include Goldman Sachs, Siemens, Swiss International Air Lines, UBS and Yahoo! As entrepreneur, he co-founded MoodLogic, voted "best music organizer" by CNET in 2003. He also advises several startups. He received his Ph.D. in physics from Stanford University in 1991 and worked as a researcher at Xerox PARC (Palo Alto Research Center). He lives in San Francisco, Shanghai and on **weigend.com**.

BIO **Jochen Wegner** is deputy science editor of *Focus* magazine, regularly publishing cover stories about psychology, high tech, risk research and medicine. Wegner also writes books, including the best-seller in Germany about chance, *Why me? – Fate. A User's Guide.*

Find out about SAS data mining solutions**: www.sas.com/datamining**