Netting data

Adept computer scientist and journal editor **Dr Jim Jansen** gives an insight into his career, explaining not only the key parameters of his current research projects, but also his teaching methods and the challenges associated with bringing articles to the market quickly

You are currently a professor at Pennsylvania State University's College of Information Sciences and Technology, a scientist with the Qatar Computing Research Institute and have held a variety of other roles relating to research and publishing. How did you first come to be involved in computer science?

Before I came into academia, I was in the US Army, where I was a communications officer. As part of that role, I was asked to teach at the US's military academy, West Point, in the Computer Science Department. This started my graduate school experience, which eventually led me to get a PhD in computer science. After I left the military, I went immediately into academia – where I found a very similar organizational structure, only without uniforms!

Your current research goal is to increase effectiveness for accomplishing information tasks by improving interactions among people, technology and information. What motivated this line of enquiry?

I particularly enjoy applied research, and when I was in graduate school the Web, and the Internet upon which it was being built, were just entering the mainstream. There was a critical need at that time to access information and use information technologies in a way to accomplish tasks, achieve goals and entertain. Even now, with some of the most successful information engines humankind has ever created (such as Google), we are still grappling with this problem.

The main focus of your research has been around Web use and relationships. What have been some of your most recent findings in this area?

It is difficult for most of us to imagine our daily lives without the Web, or Internet and mobile technologies. The majority of my work recently has been focused on financing; folks just don't realize the massive amount of cash it takes to keep a major search engine running. There is also little financial consideration of the cost of the 'free' stuff that search engines provide. Almost all of that free stuff is paid for by online advertising. In fact, if search engines did not rely on the business model that they currently use, the Web would look very different from how it does today.

Web searching, information retrieval, keyword advertising, online marketing and social networking are some of your current areas of investigation. What are the primary methods you use to conduct your research projects?

Digital analytics. As many people will already be aware, when we make use of online technologies, we leave traces of our activity that are collected by various systems. The analysis of this trace data can tell us many things about not only individuals, but also businesses, governments and societies. Even things that we may not want to hear about!

Could you expand on your involvement with the Google Online Marketing Challenge?

The Google Online Marketing Challenge was one of the best educational endeavours ever initiated and supported by a business. Its impact on the entire online advertising industry is nearly immeasurable. I hope Google continues to give it the attention that it deserves, since the benefits – though indirect – have contributed to a much more effective and efficient advertising industry.

Do you believe that case studies, simulations and technology are important in engaging the students of today and tomorrow?

Sure – but we are an information-rich world now. The need for professors to provide information for their students in a lesson is far less pressing than it was; what is critical is the student's attention. That's why I'm a big believer in muscular learning: teaching by providing some key information, the resources to access further information and focusing on involving students in hands-on tasks and activities while in the classroom. In other words, focusing their attention instead of providing information.

You are the editor-in-chief of the journal *Internet Research*. What attracted you to work in publishing and how has your experience been to date?

I've published in and reviewed for *Internet Research* for years prior to becoming editor. I wanted – and still want – the journal to be a source for impactful and meaningful research in the Web space. To attract good researchers, however, like any other publisher, we have to get articles to market quickly – and meeting this narrow time-to-market target has been my biggest challenge to date. The reviewers, who are all volunteers, are getting reviews back in about 30 days. Now, the journal needs to get those articles online in about the same amount of time.

Online behaviour

Computer scientists at **Pennsylvania State University** in the US have, in recent years, made inroads towards better understanding the behaviour of humans online – with special regard to their interactions with search engines, social media sites and online adverts

IN THE LEXICON of advertising and Web research, the term 'digital native' is beginning to rise to popularity – reflecting the inexorable growth of the group of Internet users it represents. Although most people on the planet were not born with on-demand access to digital systems, digital natives are an exception to this rule as today such systems pervade – and have in fact become indivisible from everyday life. As an increasing number of activities are therefore accomplished through the mediation of smart devices and telecommunication technologies including the Internet, the way that people use digital tools and applications is quickly becoming increasingly important; understanding online behaviours and the trends they follow can be very valuable.

This may seem like a simple thing to determine; it is easy enough, after all, to observe the behaviour of shoppers on the high street, or track how many people are watching a certain TV broadcast. Furthermore, one might assume that the preferences consumers show offline will be mirrored in their online operations. The reality, however, is that regardless of whether they are pursuing entertainment, products or content, the Internet environment brings out very different behaviours in humans – and rather than being public and easy to observe, these interactions are concealed within the privacy of people's homes.

UNCOMMON INSIGHTS
Luckily, tools do exist to trace online behaviour – and based on the data collected by these tools, it is often possible to draw meaningful conclusions about how people conduct themselves on the Internet. In advertising and marketing, these kind of analytic data, along with data that people voluntarily contribute to the Internet through weblogs, social media sites and the comment features of commercial sites, are extremely helpful. They help point the way towards effective advertising campaigns, marketing methods that resonate with Internet users and good practice for brands and companies engaging with their prospective customers through digital media.

One investigator at the centre of a number of research projects in this area is Dr Jim Jansen of Pennsylvania State University and the Qatar Computing Research Institute. Jansen's work covers a range of topics within online behaviour, extending to information search and retrieval patterns, brand interactions on Twitter, the efficacy of keyword advertising strategies and consumption patterns in paid-for digital content. "My investigations are not classical computer science; however, the insights I have gained have fuelled algorithmic development and an understanding, especially in the ecommerce area, of the impact of algorithmic technology," he explains.

Positive brand experiences are more likely to be shared through Twitter than negative or neutral experiences

A DEARTH OF DATA

Online research of the nature pursued by Jansen and his colleagues is challenging from the point of view that it requires very large sets of real-world data comprised of millions of records – and this kind of information is very difficult to come by, especially on such a massive scale. Thankfully, opportunities do sometimes arise; in 1997, for example, Doug Cutting of the search engine Excite offered a set of more than 50,000 user queries for study. Jansen was one of the researchers eager to take Cutting up on this proposition – and when he met other researchers who had expressed similar interest via email, they began to collaborate towards making the best use of those records.

The result was one of Jansen's most influential early papers, published in 2000. This study considers a number of aspects of the data, including the search terms, and the use of logic and modifiers within a query, the queries themselves and the overall sessions undertaken by each user. The results were very telling: 67 per cent of users did not progress beyond their initial query, and 26 per cent stuck with only two or three queries. The average length of a query was 2.21 terms, and – following the trend for queries – most users (86 per cent) viewed three or fewer pages during their session. Among the most popular individual terms were connectives and sexual words, but a separation of terms by category also demonstrated that the use of geographical and economic terms were significant.

TWEET RELIEF

Not all Web data are so hard to access, however. Users of the social networking site Twitter typically create a very high number of micro-posts at a relatively fast rate, and these posts are readily visible to the public. Interestingly, in his 2009 paper *Twitter Power: Tweets as Electronic Word of Mouth*, Jansen defines the platform as a micro-blogging site rather than associating it with other social media outlets such as Facebook or Google+. Since the posts are archived and available in reverse chronological order on the user's home page, the cumulative total of previous tweets effectively forms a highly fragmented weblog.

The 2009 study examined the sentiments expressed towards certain brands by Twitter users, gathering together a dataset consisting of more than 150,000 tweets containing comments, sentiments and opinions on brands and products. They also investigated a smaller sample of 14,200 randomly selected tweets, of which around 2,700 made reference to brands and products, demonstrating the richness of the environment Twitter presents to marketing professionals. Of these tweets that make reference to brands, a further 20 per cent of users share their personal sentiments towards that brand. Positive brand experiences are more likely to be shared through Twitter than negative or neutral experiences, with more than half of all the tweets in this category containing the brand or product along with positive words.

KEY CAMPAIGNS

More recently, Jansen and his collaborators have focused on a particularly lucrative area of people's online behaviour: their engagement with keyword advertising. This form of advertising relies on the keywords entered by the user – usually into a search engine – to match that user with targeted advertising material. Google's AdWords service has become so prevalent in this function that it is almost synonymous with keyword advertising. It is partly for this reason that Jansen has taken an active role in helping students at Pennsylvania State University's College of Information Sciences and Technology to excel in the Google Online Marketing Challenge. This educational initiative encourages participants to spend a US $250 advertising budget through Google AdWords to develop online advertising campaigns for a local business.

Jansen's own work on keyword advertising has generated some significant insights – perhaps most interestingly concerning the subject of gender targeting in keyword advertising. During work on this topic, the Pennsylvania researchers discovered that personalizing key phrases by gender is actually counterproductive; their study showed that gender neutral phrases generated a return-on-investment 20 times higher than gendered phrases. In a 2013 study, Jansen and his colleagues also investigated the impact of ad rankings on the efficacy of keyword adverts, finding that in almost all areas higher-ranked ads performed better. Conversion rates, however – the rate at which an advert 'converts' hits into a sale – was uniform across all but the two top-ranked adverts, which achieved more highly.

THE FUTURE OF COMMERCE

High-street commerce – and subsequently advertising and marketing – is governed by complex rules derived from the behaviour of consumers over centuries of human history. As traders adapt to the online environment, these rules will inevitably be remade and redefined; it is research projects like those pursued by Jansen that will help companies to fully capitalize on the potential of the digital marketplace.

JIM JANSEN is a graduate of West Point and has a PhD in computer science from Texas A&M University, along with Master's degrees from Texas A&M (computer science) and Troy State (international relations). Jansen is editor-in-chief of the journal *Internet Research* (published by Emerald), a member of the editorial boards of seven international journals and has served on the research committee for the Search Engine Marketing Professional Organization (SEMPO).