What are your web data good for — Time for a rethink on web analytics standards

Paul Cook

Received (in revised form): 16 April 2004

Paul Cook
founded RedEye in 1997. Previously he had worked as Head of Direct Sales at Piccadilly Radio and then as Head of Sales at Emap Internet. During this time Paul identified an opportunity to apply traditional marketing principles to the developing new media environment, and provide marketers with the accountable and accurate ROI data they needed. Under Paul's guidance, RedEye has evolved into one of the industry's leading e-CRM specialists, providing software and services to help e-businesses get more customers spending more money, more often.

Keywords: web measurement, online management information, IP- and cookie-based tracking, web analytics, data-driven direct marketing

Abstract
When it comes to gathering data about website visitors, page views and activity on your site, the chances are that your business is using either a cookie- or IP address-based approach. These data are your online management information and are likely to form the basis of decisions about online business expenditure, the value of online marketing activity and the impact the web is having on your overall customer value. But what if the metrics on which you are basing these strategic business decisions are fundamentally flawed? This paper discusses the results of a study undertaken to examine this issue in more detail.

Introduction

Cookie-based web tracking solutions have typically been used by businesses during the last ten years in an attempt to obtain accurate customer data and build more profitable customer relationships.

Over much of this period the author's company has been able to track the suitability of cookies for monitoring website visitors over the longer term. In the past few years there has been something of a consumer backlash against cookies and a fair number of internet users now delete them on a regular basis. As a result, cookie data can no longer be viewed as a solid foundation for accurate data-driven direct marketing.

If cookies have their limitations then it is quite alarming to consider how many businesses are using an even older generation of web analysis technology, namely IP-based server logs. Many experienced marketers are aware of the flaws in IP-based metrics (more of which later), yet even today these figures form the basis of countless business decisions.

This study began as a response to client demands for more robust information about customer loyalty and the growing need for completely accurate customer profiles to drive one-to-one marketing activities. In what is believed to be the first study of its kind, the author's company attempted to quantify the level of inaccuracy inherent in much online management information.

The study

The purpose of this study was to investigate the accuracy of IP and cookie-based web analytics in both absolute terms and as a 'common currency'. The aim was to find out how appropriate each type of data is

Paul Cook
RedEye International Ltd
24–28 Nelson's Row
Clapham
London SW4 7JT, UK
Tel: +44 (0)20 7627 9300
fax: +44 (0)20 7627 9301
E-mail: paul.cook@redeye.com

for making business decisions, and to discover whether either approach could provide a method for comparing the popularity of different websites for media planning purposes.

The study examined two of the UK's busiest e-commerce websites, www.williamhill.co.uk and www.asda.com, over a period of 28 days. These websites were chosen because:

— there is a high propensity for customers to make multiple repeat purchases in a 28-day period
— logged-in users account for more than half of all page impressions on the sites
— the two sites appeal to very different types of consumers.

Robust sample with 100 per cent accurate information

More than half of all page requests on these websites are made by logged-in customers, providing a robust sample of known data against which to benchmark IP- and cookie-based tracking. The study used specially written software to identify people using their log-in details, thus providing virtually 100 per cent accurate information about what customers do while logged in to the site. The ability to capture a unique customer reference at the same time as both the cookie and IP address has made it possible to quantify the level of error inherent in these different approaches.

The results

IP-based tracking

The study found that the most commonly used tracking methodology, IP-based server logs, inflated website visitor numbers by up to 660 per cent over the 28-day period, which would lead to a company underestimating its conversion ratio by 7.6 times (Table 1). IP-based analysis (Table 2) proved no better at identifying the number of distinct visits to the site, over-reporting these by up to 260 per cent and only managing to identify accurately 14 per cent of visits from start to finish (ie all the pages visited in the correct order).

Cookie-based tracking

Cookie-based tracking fairly accurate over short period of time

According to the study, cookie-based tracking (Tables 3 and 4) was fairly accurate over a short period of time, but the number of people who

Table 1: Worst case IP-based metrics from two sites indexed against log-in results

IP indexed totals	1 day	7 days	28 days
Pages		100	
Total visitors	261	502	760
Repeat visitors	1,010	305	276
Total visitors		361	

Notes: 100 = 100% accurate. So for example, '760' would mean the figures were over-inflated by 7.6 times the correct figure.
All results have been indexed and combined in order to protect client confidentiality.

Table 2: Percentage of correct paths and histories using IP/ browser-based tracking

IP based	1 day (%)	7 days (%)	28 days (%)
Paths		14	
Browse history	16	14	8
Tracked completely	43	27	22

Table 3: Worst case cookie-based metrics from two sites indexed against log-in results

Cookie indexed totals	1 day	7 days	28 days
Pages		100	
Total visitors	113	158	228
Repeat visitors	90	117	155
Total visits		99	

Table 4: Percentage of correct paths and histories using cookie-based tracking

Cookie based	1 day (%)	7 days (%)	28 days (%)
Paths		93	
Browse history	72	40	22
Tracked completely	81	63	50

periodically delete cookies or use more than one computer leads to significant inaccuracies over a prolonged period. Cookie-based analysis led, over the 28-day period, to an overstatement of the total visitors figure of up to 128 per cent, with only 50 per cent of visitors tracked uniquely over that time. One implication of this is that online marketing campaigns may be twice as effective as previously thought.

Online marketing campaigns twice as effective as previously thought?

Finally, Figure 1 illustrates how quickly the inaccuracies of reidentifying visitors using an IP-based or cookie-based approach can mount up.

What causes the errors?

IP data

To start with IP-based figures, which carry the greatest error, when you log on to the internet your ISP assigns you an IP address, which is also a multi-digit number. When you request a URL your computer sends a packet containing the IP address so the server knows where to send the page. But ISPs have multiple proxy servers to handle all the traffic and your request is handed over to whichever machine is least busy (Figure 2). So it is the IP address of this machine that the web server sees and this can change for every page requested. One person looking at three pages can look like three different people looking at one page. Conversely, two people browsing from behind the same external-facing server can share the same IP address and appear as one person.

The number of users and web domains is continually going up but there

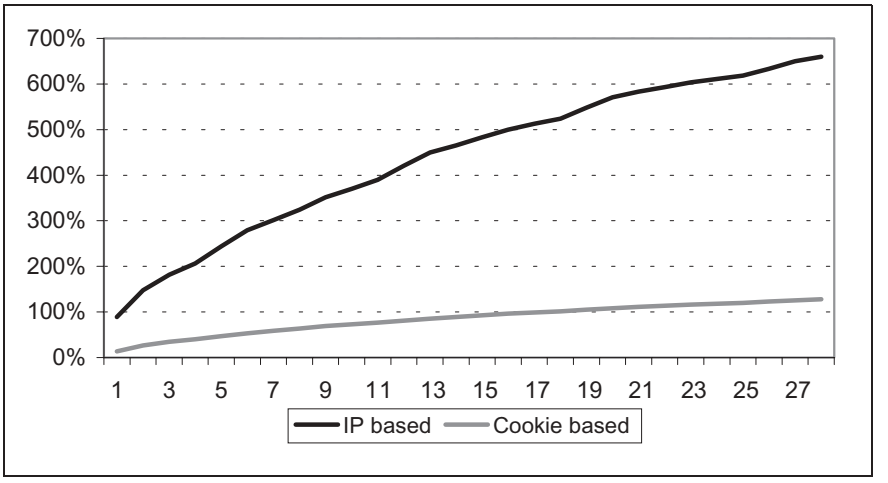

Figure 1: Percentage error in cumulative unique visitor figures over a 28-day period on one of the sites

are only a fixed number of IP addresses available, so different users frequently share the same IP address. To save costs ISPs also tend to maintain only a small number of IP addresses, which they share among their entire user base.

IP-based tracking systems can over- and under-count site visitor numbers

All these factors mean that IP-based tracking systems can over- and under-count the number of visitors to a site. IP addresses are simply incapable of accurately tracking.

Cookie data

Under a cookie-based approach the visitor is requested to accept a unique reference that they should use every time they request a page from the site. The visitor can choose whether or not to accept the cookie via the settings in their web browser.

People who delete cookies can reappear as new customers

The main issue with cookie data is that people who delete their cookies can reappear as new customers, leading to an underestimation of

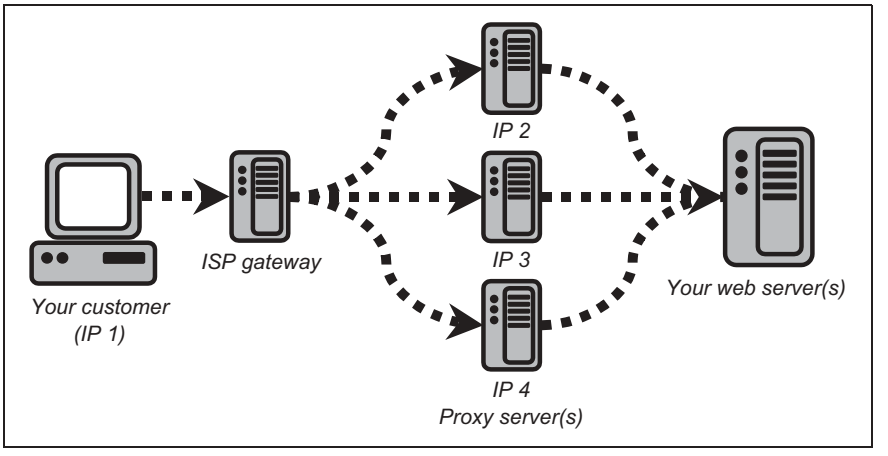

Figure 2: How the customer's true IP address is hidden behind their ISP. The same customer can appear as any of the IP2, IP3 or IP4 within a short period of time but never as IP1

customer loyalty and conversion ratios. Secondly, people often use more than one computer to access the internet, and given that each machine may well have a different cookie attached to it, they will appear as more than one person.

An NOP study[1] of 1,000 'British internet users' (definition: spend at least one hour online in an average week, representing 80 per cent of the GB internet base) offers further explanation of these factors. Fifty per cent of respondents said they had used more than one computer in the last three months, while 89 per cent of respondents deleted cookies periodically if they knew what cookies were and how to delete them.

Accuracy of cookies diminishes over time

Cookies remain a good way of tracking individuals over shorter periods and are certainly more accurate than IP addresses, but their accuracy diminishes significantly over time.

Reflections

These results will be particularly alarming for companies which rely on IP-based software to analyse the effectiveness of their online activities. The inability of IP-based software to track visitors for even 30 minutes (the length of a visit) means that the only figure these systems can provide reliably is the total number of times a page was accessed.

IP-based metrics: little use for those serious about accuracy of their business/marketing data

On a personal level, these findings cement the author's long-held belief that IP-based metrics have very little use for anyone who is serious about the accuracy of their business and marketing data. Despite this, many businesses continue to make multi-million-pound decisions on the strength of what is essentially bargain-basement management information.

This study also confirms that the accuracy of cookie data diminishes significantly over time. People who delete cookies can appear on subsequent visits to be new customers, potentially leading to an underestimation of customer loyalty and conversion ratios.

Cookies more accurate than IP addresses

Clearly cookies provide a much more accurate picture than IP addresses, and are actually quite effective for tracking individuals over a 24-hour period. But analysis of customer data over a longer period should be carried out with caution. Most of the data that marketers need can be obtained by weighting figures appropriately. Only correctly weighted cookie data will deliver good-quality management information on which to base business decisions.

So what are your data good for?

Having established that your metrics may not be as accurate as you thought, it is vital to understand what your data can and cannot be used for (Figure 3). Marketers may find the following guidelines useful when attempting to interpret management information.

IP-based data

An IP-based tracking solution is useful for tracking page impressions and banner ad impressions.

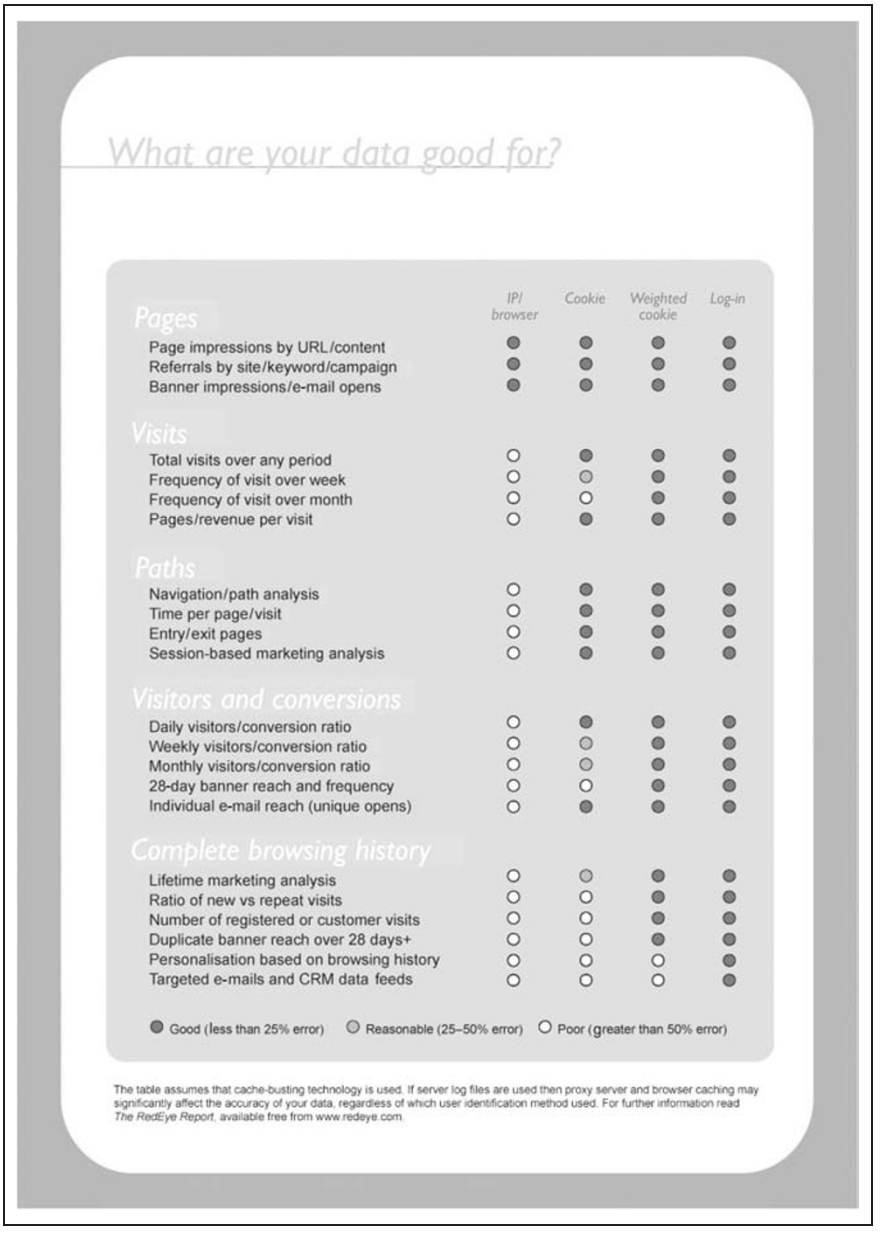

Figure 3: What you can and cannot do with your current web data

Cookie-based data

A cookie-based tracking solution can help you understand site usage and usability, including metrics such as frequency and length of visit, pages per visit and revenue per visit.

Cookie data can also be used to conduct path analysis, monitor entry and exit pages and perform session-based marketing ROI analysis.

Weighted cookie data

As we learn more about the errors inherent in cookie data over longer periods of time, it should become easier to weight the data to

counterbalance these errors. A cookie-based tracking system with weightings over a longer period should therefore be capable of accurately measuring:

— conversion ratio
— revenue per visitor
— new/repeat visitors
— banner reach/frequency.

Log-in data

Log-in data: the Holy Grail for marketers

The Holy Grail for marketers is for their customers to log in to the website. A log-in-based tracking system, as used by William Hill and Asda, opens the door to the following metrics and activities:

— complete browsing history
— lifetime marketing ROI analysis
— registered visits/visitors
— customer visits/visitors
— e-CRM e-mails based on past browsing and purchasing.

Conclusion

People have long suspected the inadequacies of IP/browser-based systems, but few could have realised the full limitations of a technique that was even used to value 'dot.com' businesses during the internet boom of the late 1990s. Clearly, as a marketing and business aid, IP-based data are of little use and the time has come to stop using bargain-basement systems to support multi-million pound business decisions.

Overall, the study offers mixed news for marketers. On the one hand, you are likely to have far fewer people visiting your site than you previously thought. But on the other hand, both the repeat visit rate and conversion ratio are probably significantly better than your management information is currently suggesting.

There is mixed news for online advertisers too. Campaigns may not be reaching as many people as they have previously believed, but the problems with using cookies to track responses over a 28-day period means their return on investment may be as much as double what they think it is.

Companies looking to personalise their online customer communications using cookie-based data should pay particular attention to these results. The ability to track consumers using only cookies is almost certain to degrade over time, as consumers become more tech-savvy and delete their cookies. Can marketers really be sure that they are sending send the right message to the right person at the right time if they only know 50 per cent of what has happened in the last 28 days?

By far the best option for successful data-driven marketing is to identify visitors via log-in wherever possible. But the author's company is also working to develop guidelines for the weighting of cookie-based management information across different types of website.

No one can claim to have all the answers, but we cannot afford simply

to brush the issue under the carpet. As the internet evolves we have the opportunity to create some truly effective web measurement standards — no other medium is more measurable.

Reference

1. Unpublished research commissioned by the author's firm.