

CRM Data Mining – Web Traffic Analysis

Finding Real Data that You Can Act Upon ...

Tina Reiners, HarrisonGray Partner

An Introduction to Web Traffic Analysis

The web traffic analysis market barely existed a few short years ago and is already a multi-million dollar segment of the e-business industry. Web traffic analysis growth is being driven by the growth of the world wide web and the desire to know as much as possible about visitors, through self-identification, registration, and web server logs. According to a recent report by International Data Corporation, the web traffic analysis market will break 100 million dollars by the year 2002.

Web traffic analysis tools take web server traffic information and try to make sense of it so intelligent business conclusions can be drawn. Simple things like how many total files were requested can be easily calculated and reported. By looking for multiple requests from the same computer during the same timeframe, more complex things can be calculated, like the number of total visitors and visits that were made to a site. By adding other information to the analysis, such as advertising information, ad impressions and click through rates can also be calculated.

Two types of web traffic analysis are described below, namely web log analysis and web mining.

Web Log Analysis – Traditional Web Traffic Analysis

Web log analysis software reports basic traffic information based on web server log files. Tools in this category use calculations and assumptions to create a maximum amount of log data relationships for inclusion in reports.

The main purpose for web log analysis has traditionally been to gain a general understanding of what is happening on the site. Webmasters and System Administrators who are responsible for keeping the site up and running, often want to know how much traffic they are getting, how many requests fail, and what kinds of errors are being generated. This information is typically used for web site management purposes.

Recently web log analysis has become more popular with web marketers. By adding information such as advertisement names, filters, and virtual server information, log data can be further analyzed to track the results of specific marketing campaigns. Product Managers and Marketers who are responsible for allocating budgets in the most efficient manner require this type of information to make intelligent business decisions. Web log analysis can be used to answer questions like:

- What companies are visiting your site?
- What pages are the most and least popular
- What sites are your visitors coming from?
- How much bandwidth does your site utilize?

Web Mining – Advanced Web Traffic Analysis

Rather than look at web traffic data as its own island of information, web mining integrates web traffic information with other databases in the corporation such as customer, accounting, profile, and e-commerce databases. The resulting reports not only use advanced relationships between log data, but also draw from these external databases as well.

“Web mining enables discovery of meaningful business correlations and trends by integrating and analyzing Internet/intranet web traffic information and traditional business data.”

The main purpose of web mining is to analyze online investments of the entire enterprise, in an effort to maximize return. Executive management and Chief Information Officers are typical candidates for this type of information.

Many Web Miners base their offers on visitor profiles and, more importantly, create new products that match the results of their analysis. Web mining is typically used to answer more complex web-related questions like:

- How do visitors’ demographic and psychographic information correlate with their web site browsing behavior?
- What is your web site’s return on investment?
- Which advertising banners are bringing the most qualified visitors to your site?
- Which sites refer the highest number of visitors who actually purchase?

The Need for Web Traffic Analysis Tools

It is clear that a lot of time, attention, and money has been spent on the web by companies wanting to get involved in cyberspace. However, very few of them have much of a feel for their payback on this investment. Much of this is due to the incredible hype and fast growth surrounding this technology, combined with the low cost of experimentation. All one really has to do is invest in a web site and revenues will shoot through the roof, right? Well, not exactly...

It is important to realize that the web is fundamentally different than other marketing venues:

1. The web is anonymous. In traditional marketing, it's relatively easy to construct a profile of your target audience in terms of demographics and psychographics. With the web, however, you don't know much at all about the visitors to your web site – as most visitors are anonymous. They come and go from your site without much of a trace – other than the general information collected by your web server.
2. The web is interactive. Many companies overlook the fact that the web is not just a transmitter of information, it is also a receiver. At a minimum, what is received is information on what pages are being looked at by what visitors. More sophisticated sites also collect demographic and psychographic information from visitors as they browse, forming a marketing database in the process.

The ultimate goal of web traffic analysis is to combine both anonymous web traffic information with traditional demographic and psychographic business information. This allows measurement of what people say, how they feel, and most importantly, how they actually respond. This information is the foundation of personalized one-to-one marketing techniques, allowing a business to target specific audiences with customized products and services that directly solve their problems.

An Overview of Web Traffic Analysis Software

To accomplish its goal, web traffic analysis software must be able to collect web traffic data from multiple web sites, store it into a data warehouse, integrate it with other sources of information, and then analyze and report the results quickly and accurately. The ideal web traffic analysis system is also both completely programmable and extensible to support customization and scalability with the enterprise. This section goes through an overview of the technology typically used to accomplish each of these tasks. Good web traffic analysis software is available today and ranges in price from free to over \$20,000. The most popular packages are currently priced at about \$300.

Data collection

Web traffic analysis tools must first be able to collect web traffic information, often from multiple web sites deployed throughout the world. This can be accomplished through reading web server log files or more recently, through TCP/IP packet sniffing techniques.

Web Server Log Files

Web server log files are undoubtedly the most common source for web traffic information. Behind every web site is a web server, whose purpose in life is to respond to visitors by locating and sending the requested files. After each request, the web server logs the results of the exchange in a “log file”. A typical log file is ASCII based and contains information about which computer made the request, for which file, and on which date.

These log files contain useful web traffic information. By looking for multiple requests from the same computer during the same time frame, conclusions can be drawn about the total number of visitors and visits that were made to a site.

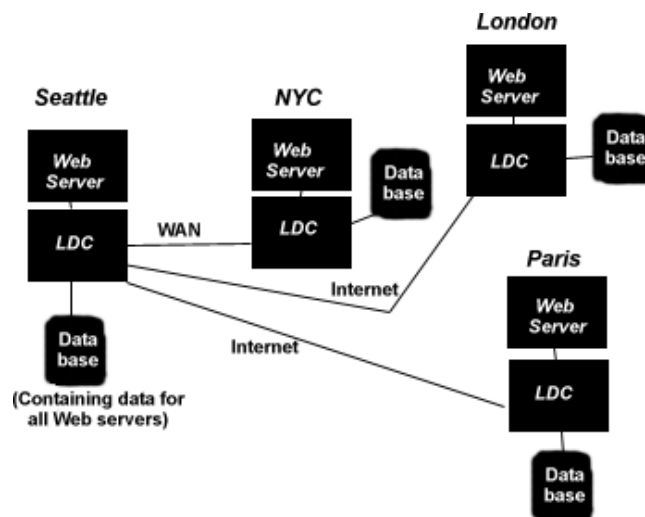
Data Collection through Packet Sniffing Technology

Packet sniffing technology has recently been introduced into the traffic analysis market which eliminates the need to collect and centralize log file data entirely. This technology gets its information directly from the TCP/IP packets that are sent to and from the web server.

The advantage here is:

- Data is collected in real time, rather than being read in from a log file after the fact. This keeps the data warehouse up to date on a continuous basis.
- Data is continuously being read into the data warehouse, rather than being collected from huge log files. This increases the data warehouse capacity.
- Companies with distributed web servers can easily and automatically collect information in a centralized data warehouse. This solves the problem of collecting all the latest log files from sites located throughout the world.

Packet sniffing technology watches network traffic going to and from the web server and extracts information directly from the TCP/IP packets. The data collector must be installed on a computer located on the same network segment as the web server that it is supposed to monitor, in order to “see” the network traffic as it goes by. Most packet sniffing tools are priced over \$10,000.



Packet sniffing allows for multiple distributed web servers to send traffic information over the internet to a centralized data warehouse. Here, Live Data Collectors (LDCs) are used in New York, London, and Paris to form a data warehouse in Seattle.

Data Integration

For web mining applications, web traffic data must be linked to other traditional business and marketing databases within the company. These databases might include e-commerce, profile, accounting, and customer registration databases for example.

A typical way to perform this is through the computer IP address, as all requests to your web server will include this information. Once you link an IP address to a particular company or person, you will be able to correlate future visits from the same IP address. This method is not perfect, however, as it is common for IP addresses to be shared among multiple users. This is done by larger organizations, such as Microsoft, as well as by Internet Service Providers such as America Online.

A more accurate method of creating this link is through the use of "cookie" technology. This ensures that the same computer is connected to your site, independent of the IP address that is used to make the connection. Another common field to link to is the "authenticated user" field.

Data Reporting and Analysis

This is a typical area that web traffic analysis tools overlook. It is important to keep in mind that the end goal of web traffic analysis is to allow for on-line investments to be quickly analyzed and business decisions to be made. These reports therefore must include real data that you can act on, rather than just reams of detailed technical information. If you can't easily analyze this information, put it in a presentation quality format, and quickly get it to the right decision maker, there isn't much purpose to collecting the information in the first place.

A quality web traffic analysis package will include features like multi-level filtering, remote reporting, and will support multiple output formats. These features make web traffic information easy to obtain for everyone in the organization.

Extensible and Programmable

The final essential feature for web traffic analysis tools is the ability to easily program the tool to integrate it with other web applications you may have, or extend its capabilities. This is a key requirement for any organization that has had to wrestle with a tool that couldn't grow with the organization's needs and eventually needed to be replaced, an extremely painful and expensive process.

The ability to program a tool, through a standard scripting language, makes it easy to add custom functions. This might be used, for example, to remotely administer advertisements through user created ASP pages.

Beyond programmability, the ability to add plug-ins is a requirement if you need to purchase or integrate your own custom functions. This can be used to extend data collection capabilities to other file types such as advertising servers and streaming-media logs, for example. It can also be used to create custom output types that might be necessary for proprietary data analysis systems.

Some Important Web Traffic Analysis Concepts

To better understand web traffic analysis software and what visitor information can be obtained from a web site server, it is useful to have a basic understanding of how a web server collects and logs visitor data.

Here are some key definitions that are imperative to understand when analyzing web sites:

Hit or Request - A "Hit" or "Request" refers to an individual file request made to the webserver. This can be measured very precisely by simply counting the number of lines contained in the web server log file. Measuring requests, however, is not a very accurate measure of website popularity, as each visit to a site can generate large numbers of file requests. The way a web site is designed and the number of graphics it has will both significantly effect the number of file requests a site receives. In the example above, there are 26 requests.

Page View - A better measurement of site traffic can be found by counting page views. A page view is simply the transfer of a specific HTML file. Page views can also be measured precisely by simply counting the number of requests for HTML files. Page Views are a better measurement of web site popularity, but are still imprecise when multiple HTML files are required to display a page (when using frames on a site, for example).

Visitor - A visitor is defined simply as a unique computer "IP address". This measurement is less precise due to the fact that IP addresses do not always have a one-to-one correlation with computers. Visitors will be **UNDERREPORTED** in the case of proxy servers (many people using the same IP address) and **OVERREPORTED** in the case of on-line service providers (one person using multiple IP addresses). This precision can be improved by the use of a persistent cookie (more on this technology later). Visitors are a much better measurement of gross web site traffic than page views or requests. In the above example, there is only 1 visitor.

Visit - A visit is calculated by grouping requests from the same IP address, the same browser type, in the same time period together (we assume that the IP address and browser don't change during a visit). The total number of visits is usually more than the total number of visitors because each visitor can visit the site more than once. Compared to "Hits" measuring visits is a more accurate measurement of gross web site popularity, but is also less precise because there is no way to be certain that a series of requests actually belongs to the same person. Measuring visits is also a good measure of gross web site traffic popularity. In the example above, there is only 1 visit.

The above measurements highlight the fact that measuring web site popularity is an imperfect science. The art of good web traffic analysis includes the ability to draw business conclusions using data that is imperfect. Having an understanding of the data, how it is collected, and what limitations exist is key to drawing the proper business conclusions. As demonstrated in our example, one must understand and trade off between **PRECISION** (Requests and Page Views are precise but less accurate) and **ACCURACY** (Visits and Visitors are more accurate, but less precise) before drawing conclusions.

Getting Fancy with Web Traffic Information

There is actually more information than most people realize in a log file (see Appendix A). With this information alone, the following can be easily calculated:

- How many Hits, Page Views, Visits, and Visitors am I getting? (as discussed above)
- Which browsers and operating systems are the most common? (through the agent field)
- How many bytes are being transferred by the server? (through the bytes sent field)
- Which files and directories are the most popular? (through the requested file names)
- What sites are visitors coming from? (through the referrer field)

But wait! With a little more work, there's still more information we can add to this analysis. Log file analysis is a lot like music composition - by combining the basic information (the notes) in creative ways (the music), we can extend the analysis and draw meaningful conclusions.

Converting IP Addresses back to Domain Name

Every visitor to your web site connects to the internet through an IP address (i.e. 193.237.55.144). Every IP address has a corresponding domain name associated with it and these are linked through the Domain Name System (DNS). When a visitor enters a website address into their browser, it is the DNS system that converts this name into the

appropriate IP address (206.129.192.10) so that the computers can connect with each other. We can use the DNS system in reverse (called a reverse DNS lookup) to convert visitor's IP addresses back into domain names.

You can either have your web server perform this DNS conversion for you, or have your web traffic analysis software do it. Having your server do it will slow your web server down. In addition, if someone else hosts your site, you may not have a choice in the matter, as most service providers don't want to slow their servers down to perform reverse DNS lookups.

Converting File Names to Page Title

A well designed site will have a title (using the TITLE HTML keyword) for every page on the site. Rather than simply report the file names that were requested, we can easily look at these files and determine the corresponding page names. In general, page names are much more human friendly in terms of communicating information. By simply extracting the page names from the files that are listed in our web server logs, we can end up with a report that contains actual page names, rather than simply the file names themselves.

Path Analysis

By linking log file entries and then sorting by time and date, we can also start to see the path that a particular visitor took through the site. If we do this for each individual visitor, we are able to calculate the most popular paths taken through the site as a whole.

Grouping information

By grouping information together, we can start to draw higher level conclusions. For example, if we group all Netscape Browser and all Microsoft Browser information together we could calculate which company's browser was the more popular on our site. Using this same technique with referrer information, and looking for any referring URL with the word "Yahoo" in it, we could see how many of our visitors came from Yahoo as a whole.

Countries

By looking at the extensions on our visitors domain names, we can also estimate where in the world our visitors are coming from. Example extensions include:

- .ca - Canada
- .au - Australia
- .se - Sweden
- .uk - United Kingdom

Filtering information

By filtering information, we can answer very specific questions about the site. For example, to calculate how many visitors we got from Microsoft this week, we would only look at information from this week, and only look at visitors that have the word "Microsoft" contained in their domain name. This could be compared to our overall traffic to determine what percentage of our visitors presumably work for Microsoft.

Correlating information

By correlating and cross-tabbing information, we can answer questions like "Of the visitors I get from Germany, how many of them use Microsoft Windows 98 as their operating system?" This kind of detailed information can be useful for both site management and marketing segmentation purposes.

Query string parsing

Query strings are typically used on database driven sites and consist of all of the data at the end of a URL (usually delimited with a "?"). For example, the following referring URL is from Yahoo:

<http://search.yahoo.com/bin/search?p=Web+Traffic+Analysis>

By looking at the data after the "?" we see that this visitor searched for "Web Traffic Analysis" on Yahoo before coming to our site. Yahoo encodes this information with a query parameter called "p" and separates each search keyword with the "+" character. In this example, "p" is called the query parameter and "Web", "Traffic", and "Analysis" are each referred to as parameter values.

This information would normally be stored via the query field in the log file and by looking at it in detail we can draw conclusions about what our visitors were searching for before hitting our site. This is useful information when trying to design the site to come up higher on the list of search engines, or when trying to determine what visitors are looking for before coming to our site.

Virtual Servers

If you host multiple sites, another useful piece of information is contained in the site name field. By using this information intelligently, we could perform a separate analysis for each of the sites that you host. This would tell you which of your sites is responsible for the most traffic.

Adding cookies into the picture

Another field contained in the log file is the cookie field, which is a topic that has received much attention and debate in the press. Up to this point, we have really only been discussing the tracking of IP addresses as they come to our site. Cookies were invented to attempt to do a better job of tracking people, rather than simply IP addresses.

This technology was developed by Netscape and is really pretty ingenious. A cookie is merely a unique identifying code that the web server gives to the browser to store locally on the hard drive during the first visit to the site. The intent of a cookie is to uniquely identify visitors as they come to the site.

Cookies benefit web site developers by making individual "requests" much more trackable, which results in a greater understanding on how the site is being used and therefore a better web site design. Cookies also benefit visitors by allowing web site to "recognize" repeat visitors. For example, Amazon.com uses cookies to enable their "one-click" book ordering. Since the company already has your mailing address and credit card on file, they don't make you re-enter all of this information to complete the transaction. It is important to note that the cookie did not obtain this mailing or credit card information. This information was collected in some other way, typically by the visitor entering it directly into a form contained on the site. The cookie merely confirms that the same computer is back during the next visit to the site.

Unfortunately, cookies remain a misunderstood and controversial topic. Contrary to many beliefs, a cookie is not an executable program, so it can't format your hard drive or steal private information from you (note that languages like Java CAN do either of these things, but for some reason Java doesn't get the negative "security" press that cookies do). The second objection regarding cookies is that some people feel that it is a violation of their resources to be forced to store information on their computer for the benefit of the web site's owner. In reality, the amount of disk space that a cookie takes up is trivial. Regardless of how you feel about cookies, modern browsers all have the ability to turn this feature off, and not accept cookies. If your site uses cookies, this information will show up in the cookie field of the log file and can be used by your web traffic analysis software to do a better job of tracking repeat visitors.

Web Mining - Going beyond web server log files

Web mining technology allows you to extend your analysis even further, by incorporating other information along with web server log files into your analysis. This allows for information to be correlated to web browsing behavior, such as accounting, profile, demographic, and psychographic information. Complex questions like the following can therefore be addressed:

- Of the people that hit our web site, how many purchased something?
- Which advertising campaigns resulted in the most purchases (not just "hits")?
- Do my web visitors fit a certain profile? Can I use this for segmenting my market?

Measuring Return of Online Advertising Campaigns

As online advertising banners become more popular, companies using them must accurately measure overall return on advertising investment. Reporting advertising effectiveness is beneficial not only to advertisers, but also to the sites and companies hosting the ads, allowing them to adjust proper advertising rates accordingly.

Proper measurement of advertising reports on two specific areas:

1. **Quantity:** How many impressions were delivered for each ad banner and page, and how many people clicked on each ad? These are usually reported as impressions and click-throughs.
2. **Quality:** Of people who clicked on an ad banner on a specific page, how many actually purchased? This return is best measured by taking into account the advertising expenses and the resulting revenues.

For companies who offer ad space on their site, reporting ad impressions and click-through rates for any page running advertisements is important. For companies running banner ads, measurement of prospect quality can also be delivered. Generally, it is nice to isolate not only which ad banners brought the most qualified visitors, but also which pages that the ad banners ran on, which also provides the most qualified users. By addressing this, an advertiser can place the most effective ad banner on the most effective page.

The following report gives an example of an online automobile sales site and the most effective ads for each page.

Highest Click-through Rates for Each Page					
Page Name	Ad Name	Impressions	Click throughs	Click through rate	Cost
Front Page /default.htm	Mustang	34100	2100	6.2%	\$3410
	Sebring	34600	1400	4.0%	\$3460
	Corvette	92100	3100	3.4%	\$9210
	Intrigue	64100	2100	3.3%	\$6410
	Camaro	93700	1500	1.6%	\$9370
Classifieds /class.html	Corvette	9800	300	3.1%	\$980
	Intrigue	10000	200	2.0%	\$1000
Hot Topics /hotnews.asp	Mustang	3400	1200	35.3%	\$340
	Corvette	3300	200	6.1%	\$330
	Sebring	5900	200	3.4%	\$590
	Camaro	6200	200	3.2%	\$620
	Intrigue	6800	200	2.9%	\$680

Measuring Return of Email Campaign

One of the best ways to measure email campaigns involves using an email merging program to send custom URLs with unique query strings to each prospect. For example, a link might be included in the email message that contains the prospects' email address. This might look like:

http://www.company.com/default.htm?Visitor=email_address

When this was clicked on by the prospect, the unique identifier (email_address) would be passed back to the company. By using a filter based on parsing the query string, it is not only possible to measure how many prospects are visiting the site, but also who they are, thus linking them back to the original marketing database. With DataLink, it would then be easy to track the specific results, in dollars, for each email campaign.

Measuring Expense Saving

A web site investment can deliver ROI benefits through cost savings, as well as revenue generation. Every visit that gets a question answered via the web site, as opposed to calling the company, represents a cost savings through reduced technical support and literature development/ mailing costs. Web mining and traffic analysis can help determine what people are looking for on the site and whether they actually got their questions answered. By combining this information with an expense database (or spreadsheet), an estimate of cost savings could be generated each month based on specific visitor requests to the web site. In addition, this information could be used to improve the web content, expanding on the most popular search topics.

Qualifying and Distributing Leads

Web mining can be used to not only collect leads, but also qualify and distribute them to your sales force as well. For example, imagine using the Dun and Bradstreet SIC database to integrate corporate information along with your web traffic information. Leads could be sorted by territory, then by company size and distributed worldwide to the proper sales territory. This process can be completely automated, distributing web site sales leads as often as you would like.

The following report shows an example of distributing domestic leads for a large U.S. based software company to the Vice President of Domestic Sales. A similar report is sent to other territories, along with a database format for importing into the local database. This process is completely automated on a weekly basis.

Top Sales Prospects (Sample Analysis for a VP of Sales)						
Shows a list of the most common U.S. visitors to the site. This information has been generated by matching web traffic data stored in the Hit List database with information stored in the Dun and Bradstreet database.						
Company	Contact	Email	Telephone	SIC Description	Annual Revenue	Visits
Microsoft	Seth Longo	sethl@microsoft.com	206-962-1200	Software	1 Billion	65
Tango Designs	Suzanne Gayaldo	sgayaldo@dirk.tango.net	509-323-6027	Computer Design	20 Million	64
Intel	Sanford Arnold	sa@intel.com	212-865-8584	Computer Hardware	1 Billion	60
Lucent	Larry Rubin	larry_rubin@lucent.com	201-386-4200	Telecommunications	950 Million	49
Volvo	Yasim Kinneer	ykinner@vd.volvo.se	206-765-1008	Auto Manufacturing	900 Million	48
ESPN	Michael Louis	orioles@espn.com	203-585-2000	Entertainment	300 Million	46
USDA	Tom Bianchi	tommyb@dc.usda.gov	202-548-2435	Insurance	40 Million	36
Apple	Mark Sojic	mark@apple.com	310-255-7500	Software	50 Million	36

In addition, this data can be used in a marketing report showing what pages are the most popular by industry (SIC code). This information could then be used to better target site information to particular industries.

Most Popular Pages By SIC Code		
Shows which pages are most commonly requested by companies with the following SIC codes and SIC descriptions.		
Page Name	SIC Code	Total Requests
Home Page.htm	3454 - Manufacturing	4,090
	5466 - Software	3,000
	8745 - Real Estate	3,500
Pricing Page.htm	0343 - Construction	10,057
	2354 - Insurance	1,300
Ordering Page.htm	6404 – Banking	3,700
	2111 – Manufacturing	2,300
Reselle Page.htm	9999 – Financial Services	3,516
	6854 – Telecommunication	5,400

Performing Marketing Segmentation

When combined with a profiling system, web mining can be used to perform marketing segmentation. This allows web marketers to better target campaigns and messages to each target group.

For example, an online music company using a profiling system could easily create reports detailing the differences in browsing behavior based on age ranges. They might find that most of their actual purchasers are in their twenties. An understanding of what information was attractive to other visitors would be invaluable in designing the web site to appeal to a wider audience. This information could then be used to expand content and quickly direct visitors to the right place.

Here is an example report showing browsing profiles by age range.

Age Ranges per Page		
Shows the age ranges, in 10 year increments, of visitors to each page.		
Page Name	Age Range	Requests
Home Page	0: 9	1
	10: 19	3
	20: 29	23
	30: 39	13
	40: 49	11
	50: 59	8
	60: 69	3
Product Page	0: 9	1
	10: 19	4
	20: 29	263
	30: 39	141
	40: 49	23
	50: 59	71
	60: 69	19
Customer Support Page	20: 29	21
	30: 39	14
	40: 49	8
	50: 59	8
	60: 69	6
Download Page	30: 39	2
FAQ Page	20: 29	2
	30: 39	2
Company Info Page	20: 29	29
	30: 39	17
	50: 59	5
	60: 69	1

Some Comments on Privacy

Once visitors realize that their behavior on the web can be tracked and analyzed, there are often concerns regarding privacy issues. It is important to note that there is no way to obtain private demographic information from either web mining or traffic analysis software alone. This would include a visitor's name, address, phone number, email address, and credit card number. In this regard, the web remains a completely anonymous venue.

On the other hand, when someone visits a web site their movements on the site are completely public. This is useful information when reported in aggregate, as it gives the webmaster direct feedback on popular content. It also benefits the user by giving the company direct feedback on how to improve the content and navigation of the site.

Many savvy marketers also try to collect demographic and psychographic information about visitors through on-line surveys and forms. It benefits the visitor by providing the company with direct customer feedback, which result in better products and web sites. This level of information is also supplied on a completely volunteer basis.

Discussions over privacy should really focus on what a company does with the data, not what data they collect. Beware of companies that are in the business of selling or supplying your data to other companies. Responsible online companies will:

- Let visitors know what data is being collected
- Let visitors know how this data will be used
- Allow visitors to refuse to supply the information
- Educate visitors on the benefits they receive from supplying personal data

Conclusion

The Internet is the most significant technology the world has seen since the computer and has the potential to revolutionize business and marketing techniques. If you are serious about your on-line investment, the first step is to get a better understanding about who is visiting your site and what they are looking for.

Knowing this information allows you to make better business decisions by catering your marketing campaign towards your best customers and delivering the information they are looking for. With today's powerful web traffic analysis software, this analysis can be performed easily and inexpensively