An Evaluative Methodology for Virtual Communities Using Web Analytics

A. D. Phippen

Network Research Group, University of Plymouth, United Kingdom
andy@jack.see.plymouth.ac.uk

Abstract

The evaluation of virtual community usage and user behaviour has its roots in social science approaches such as interview, document analysis and survey. Little evaluation is carried out using traffic or protocol analysis. Business approaches to evaluating customer/business website usage are more advanced, in particular using advanced web analytics to develop greater understanding of their customer's use of their sites. The application of such techniques to virtual communities is discussed and experimentation of such techniques on a specific virtual community project demonstrates the potential for such techniques in the evaluation of social and culture web usage.

Keywords

Ethnography, evaluation, web technologies, virtual communities

1. Introduction

The concept of a virtual, or online, community has been credited to Rheingold (1993), denoting a collective of geographically distributed individuals bound by a common interest exploiting Internet technologies to enable communication. In the last ten years, this term have been accepted to relate to any collective of individuals using Internet technologies for a common purpose conforming to defined policies (Preece et al. 2003).

The vast majority of literature related to the evaluation of behaviour and usage of virtual communities is drawn from the social sciences, in particular social anthropology and ethnography. The much-cited work of Paccagnella (1997) examined how ethnographic technique's may be applied to the evaluation of virtual communities. Since then, the terms *electronic ethnography* (Pink 2000) and *virtual ethnography* (Hine 2000) have been used to describe the ethnographic approaches to the study of the internet phenomena, and these concepts have been developed further by other authors (for example, Warren & Skerratt, 2003).

When considering these ethnographic approaches to studying and conducting research on these communities, virtually all of the techniques used in the evaluation of community behaviour and usage can be placed into two distinct groups:

1. Primary evidence from participants – interview, surveys, etc. of user groups within the community to satisfy specific research goals.

2. Document analysis – analysis of emails, newsgroups, discussion forums, etc. to identify behaviour, attitudes, etc. in the community, again to satisfy specific research goals.

The techniques are all well established in the social sciences and are certainly effective in achieving research aims. However, this paper suggests that the origins of such evaluation in the social sciences also fails to develop the *virtual* aspect of community evaluation – while the network technologies upon which these communities function are exploited to enhance established evaluative techniques (for example, the power of web-based vs. postal surveys is well researched (for example, Hewson et. al. 2003) drawing information from the network traffic and protocols has very little discussion in literature.

In this paper we propose examining another area of web site assessment generally tied to the evaluation of successful business websites, namely advanced web analytics, and determining its suitability as a non-intrusive means to the evaluate the behaviour and usage of virtual communities.

2. Web Metrics and Analytics – Measuring Web Site Success

Web analytics is an evaluative technique originating from and driven by the business world in their need to get more value out of understanding the usage of their web sites, and strategies therein. A large organisation may invest significant resource in developing what they would like to be a strong web strategy and, as with any resource investment, the organisation also needs to be able to measure the success of such strategies.

Basic web analytics takes easily obtained statistics, or metrics, in order to be able to assess web site usage. This basic information can be drawn from web logs – the raw data held on a web server relating to core statistics about each HTTP request it serves. While this information can vary, depending on the logging format, it will generally contain information such as:

- Clienthost IP address
- Username
- Log time and date
- Method passed
- Resource requested

While raw log data provides an intimidatingly overwhelming volume of technical information, with some fairly straightforward processing (generally provided by some type of reporting software) simple web metrics can be easily drawn out. The most fundamental metrics are those such as hits and page views. Arguably, one can also identify session information – the pages a single user views in a single web site visit (whether this represents a true session is something that is debated in the literature – for example, Fletcher et. al (2002)). However, basic metrics, while serving some small use in giving raw statistics, have been dismissed in some literature as unreliable and unrepresentative (for example, Kilpatrick (2002), Buresh (2003), Schmitt et al (1999), Whitecross (2002)).

As a result of the dissatisfaction with basic web metrics and log file analysis, the concept of "advanced web analytics", or eMetrics (Sterne & Cutler 2000), was developed. Advanced web analytics aims to measure and understand the relationship between the customer and the web site through a richer analysis of web traffic and related data. Aberdeen Group defines advanced web analytics as:

"Monitoring and reporting of web site usage so that enterprises can better understand the complex interactions between web site visitor actions and web site offers, as well as leverage insight to optimise the site for increased customer loyalty and sales."
(Aberdeen Group, 2000).

The important issue to note with advanced analytics is that it is not just concerned with web site statistics, but the relationship and interaction between a web site and its customers. It does not just collect website information, it uses it in conjunction with other data, such as demographics, customer profiles and subscription information. For example, while basic clickstreams will show how a specific user interacts with a web site, it offers little information on who that user is. However, if this clickstream is associated with a specific user, and if that use has profile information stored by the organisation, the value of the information begins to become far greater.

In general, the use of advanced analytics has been centred on large business web strategies – organisations that have the resources to be able to either develop their own analytic strategies or to pay for consultants to develop them. A study on the use of advanced analytics in a multi-national airline company (Sheppard et. al. 2004) examines a specific organisation's use of advanced analytics in more detail, and shows the value the organisation places in their use.

3. Applying Advanced Analytics to Social Settings

The use of web logs to draw information regarding usage and behaviour at present made a small contribution in virtual community literature. A few studies (for example, McLaughlin et. al. 1999, Nonnecke and Preece 2000, Smith 1999, Wellman et. al. 2002) examine network logs to understand basic statistics about number of logins, time and date, audience numbers, referral, etc.. Additionally, Preece (2000) discusses the value of web logs for evaluating community behaviour and identifies the need to define metrics that could be applied to social and community settings. However, there is little evidence to date to suggest that any work has gone beyond these basic measures.

It is unquestionable that web logs do hold a huge amount of raw data (see section 3.3 for example). However, the techniques that could be used to gain richness from the information to inform on how a community is maturing over time are lacking. In order to develop the evaluation of community usage beyond basic metrics we would propose that the ethos behind advanced analytics can be equally applied to social and cultural settings. If we take the basic premise from advanced analytics to be understanding the complex interactions that take place on a web site and use them to greater understand one's customers, it is entirely acceptable to apply the same to social and community web sites.

Within a community or social setting, while finance may not be the driving force, there are factors that contribute success or failure in a site. Obvious general figures might relate to the number of community members, the number of sessions each member carries out on the site, the freshness of information, etc. However, just as a commercial website may have specific goals, the same could also be said of a social website. Therefore, it is not only important to identify core measures for community sites, but also to develop evaluative methodologies in order to be able to best determine success using the data available to the researcher.

3.1. Applying the Concepts to a Suitable Setting

While this paper has proposed the concept of an evaluative strategy to be applied to virtual communities, it is necessary to test this theory. In assessing these beliefs, the evaluation of a large research project with a specific web strategy was carried out (and is continuing to be carried out). The chosen test subject involves the Mediterranean Voices project - an EU funded project that aims to:

"create a database of the oral histories and cultural practices of the Mediterranean's cosmopolitan urban neighbourhoods" (Scott, 2004)

The project as a whole draws together resource from 12 distinct locations within the Mediterranean to fulfil a number of anthropological aims in determining what is meant by culture to the people who dwell within the locations. The majority of investment in the project went into the development of resources by researchers at each of the locations. The researchers spent a great deal of time interviewing, filming and photographing their specific locale (working with senior citizen's clubs, youth groups, music and cultural associations, etc.) to capture the cultural essence of the location. This media was then transcribed and archived so that it could be classified according to:

- Location (defined below)
- Media type (film, audio, slide shows, photographs, text, Flash images)
- Themes (see below)

Additional information relating to the resource, in both English and the local language was then added to provide a wealth of descriptive information. Finally, the media was provided to wider audience through the MedVoices website (www.med-voices.org), where it was catalogued to provide a searchable database of media that could either be browsed or searched based on a researcher's needs.

While the locations were all separate with different researchers assigned to specific locales, one of the aims of the project was to encourage the investigation of cross regional interests and cultures. The main mechanism for achieving cross regional searching was to classify resources according the themes. Seven overall themes (the person, living together, work, play, worship, objects, spaces) were defined, each being further decomposed into sub themes (for example, objects' subthemes include religious objects, toys, gifts and tourist objects). Any given resource could then be assigned one or more subthemes to provide a further means of classification. The searching facilities are such that a site member could search generally, tied to a location, tied to a theme, or in any combination, to encourage people to search beyond

their specific locale to identify shared cultural perspectives beyond geographic boundaries. The multi-lingual approach descriptors related to each resource and also the site in general aim to widen participation within the site, by providing both a local feeling of ownership (by providing the resource in the local language) while at the same time removing the potential for isolation/lack of interest by also providing the same details in English.

The project is an ideal candidate for experimentation as it has some very clearly defined goals for its use. Scott (2004) defined its intention for use to be:

as a research resource;
as a pedagogic tool;
as a source of information to the interested general public.

Its also has very clearly defined intentions to be cross-locational and cross-cultural in its distributing of information and resources. It aims to promote a culture of sharing knowledge about the studied locations to develop greater understanding and empathy among people in the Mediterranean. Therefore, we can initially evaluate "success" in the MedVoices site through understanding the behaviour of its audience related to how it views resources and draws cultural information from the site.

3.2. Experimental Approach

This experiment took place during the development phase of the MedVoices website. It was not accessible to the general public at this time but was accessible to the MedVoices researchers across Europe, who had the facilities to both add their own resources to the site, and also browse and use the site as a general user. The researchers were all aware of the overall aims of the project and came from a social sciences background. Due to these constraints, we only considered the first intention of the website for evaluative purposes. The aim of the experiment was defined as:

To determine the use of the med-voices site in terms of:
location;
language;
theme.

From this aim, we hoped to gain a greater appreciation of how the site was currently being used which would indicate the relative success of the project aims to date.

The information resources used within the experiment are detailed in figure 1:

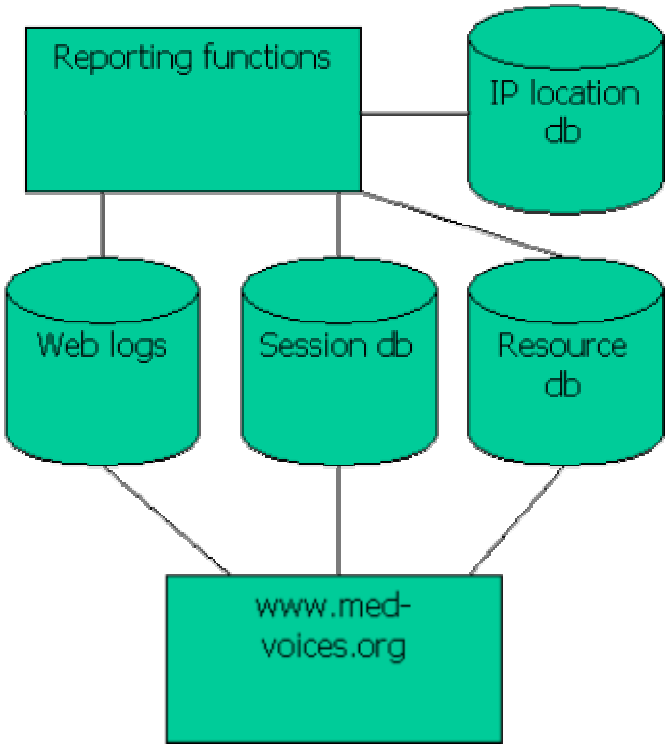

Figure 1 – Information resources in experiment

The majority of the information is drawn directly from the medvoices server. The "session db" was a small additional table added to the core resource information to hold basic session information (session identifier, client IP, time, date) to make it possible to marry session details with log information. The only other additional resource was the IP location database. In order to reduce network traffic, the GeoIP database and API (MaxMind 2004) were used to ease the resolution of IP addresses to locations. types of information drawn from these resources were guided by the experimental aims. The complete data set was loaded into a suitable relational structure, and analysed using SQL querying.

3.3. Initial Results from the Study

3.3.1. Basic Statistics

In order to establish core data, this paper details some basic measures below. While they do not help much in achieving our experimental aims, they do provide a few interesting figures when considering the volume of data available to the analyst.

From 150 Mbytes of web log data, the following basic statistics were drawn:

Hits	105604
Number of page views (*.htm* or *.asp*)	34132
Unique Client IP Addresses	318

Table 1 – Basic site statistics

While the number of unique client IP addresses obviously does not provide too much useful information, it is interesting to note that they were drawn from 87 unique class B addresses, showing a breadth of locations in a site that has ot yet been launched.

A couple of other basic statistics drawn from the logs, coupled with the session database, are:

Number of sessions	1231
No resources viewed	4672

Table 2 – Session details

3.3.2. Resource Usage

In order to understand how users of the site viewed resources, a matrix of user location vs. resource location is presented.

	Resource location					
	Valletta (Malta)	Alexandria (Egypt)	Mallorca (Spain)	London (UK)	Chania (Greece)	Nicosia (Cyprus)
Cyprus	17.28	2.47	9.88	12.35	9.88	30.86
Spain	1.4		3.15	3.5	1.4	
France	0.78	0.23	1.01	1.4	3.26	0.62
UK	4.77	2.27	0.68	9.09	18.64	0.91
Greece	0	0	0	0	59.46	0
Italy	0	0	0	0	0	0
Malta	100	0	0	0	0	0
Palestine	0	0	0	0	0	0
Turkey	1.71	0	0	6.84	4.27	0.85

	Resource location				
	Ancona (Italy)	Marseilles (France)	Granada (Spain)	Bethlehem (Palestine)	Istanbul (Turkey)
Cyprus	2.47	14.81	0	0	0
Spain	0	5.59	71	4.2	9.79
France	0.39	84.26	2.87	3.8	1.4
UK	1.35	23.64	22.27	9.55	8.18
Greece	1.35	31.08	8.11	0	0
Italy	0	40	0	0	60
Malta	0	0	0	0	0
Palestine	0	4	0	96	0
Turkey	0.21	25.64	4.91	1.71	53.85

Table 3 – % of Resources viewed based on client location

Table 3 demonstrates some very interesting results. Home locations are highlighted on the table with the greyed cells. While in almost every case the home locations show the largest proportion of resources viewed, the data also shows the level of breadth of location for resources viewed. In the majority of locations, there are resources viewed from most location. The main distinction in the data comes from resources viewed from the UK. As the

researchers in the UK are co-ordinating partners for the project, there is no surprise that they have interests in resources across the entire site.

3.3.3. Language Usage

The second measure was drawn from the viewing of resources in either the local language or English version of the resource. As there were a lot of resources only using English descriptors, in order to get a representative data set, only those resources that could be viewed in either language were sampled. Table 4 details the resource language viewing based upon location.

	Local %	English %
Cyprus	46.51	53.49
Spain	39.44	60.56
France	42.69	57.31
UK	23.63	76.37
Greece	26.67	73.33
Lebanon	15.63	84.38
Malta	33.33	66.67
Turkey	13.51	86.49
Palestine	0.00	100.00

Table 4 - % Resources view in local language and English

When considering the multi-national aims of the MedVoices project, these results are encouraging – there is a great deal of use of local language resources in the majority of locations.

3.3.4. Theme usage

In the final part of the experiment, we wished to evaluate the use of theme usage within the resource database to draw people away from their specific locale to investigate resources of similar themes in different locations. For this study, we randomly sampled a number of sessions from the database to determine whether resources viewed through themes would cause the user to deviate from their locale. In each session, we show the clickstream in terms of resource location, theme and number of resources viewed:

Session 1 (location: Spain)
Bethlehem(The Person) 2 resources → Granada (Worship) 3 resources → Bethlehem (Worship) 1 resource → Bethlehem (Spaces)

Session 2 (location: Spain)
Marseilles (Spaces) 2 resources → Bethlehem (The Person) 2 resources → Granada (The Person) 3 resource → Marseilles (The Person) 1 resource → Marseilles (Work) 4 resource → Granada (Work) 1 resource

Session 3 (location: UK)
Granada (Worship) 1 resource → Bethlehem (The Person) 2 resource → Granada (The Person) 1 resource → Granada (Play) 3 resources → Chania (Play) 1 resource → Chania (Spaces) 6 resources

Session 4 (location: France)
Marseilles (Spaces) 4 resources → Marseilles (Worship) 5 resources → Marseilles (Objects) 3 resources → Marseilles (Spaces) 3 resources

From this small sample, we cannot draw any firm conclusions. However, we can see some encouraging trends, in particular in sessions 2 and 3, where resource viewings move across both location and theme throughout the clickstream.

4. Conclusions and Further Work

The aims of the study presented in this paper were to investigate the use of advanced web analytics to virtual community settings through the study of applying such approaches to a specific virtual community. While this is only a pilot study run in the development phase of the community, the results achieved to date are very encouraging.

The implications for this study are twofold

1. The results demonstrate the value of the analysis related to the aims of the MedVoices project, and as such represents the start of ongoing work evaluating its usage. It is envisaged as the site matures, and the analytical approach gets more detailed, the understanding of the usage of the website should contribute greatly to the strategic growth of the site. The project is also developing more complex ways of relating resources together to further remove the rigid locational structure initially imposed by the website structure, and these will also contribute to further analysis.

2. The results demonstrate the effectiveness of an analytical methodology in the evaluation of a socially focussed website. While this is a single study, the potential for such techniques is encouraging. As well as further study on the MedVoices site, it is envisaged that this work will be applied to other virtual communities with which the author is involved, in order to move closer to a more general approach to applying analytic approaches to virtual community evaluation.

In developing the techniques, it is proposed that further analytics techniques are applies to the evaluative framework – in particular we are interested in applying OLAP techniques (Thomsen, 2002), to the ever-increasing volumes of data available to the evaluator. While the methods used to date have proved solid, it would be interesting to use more expressive approaches to the analysis of the data. In developing these approaches, it is hoped that such evaluative techniques will contribute to the growing need to understand virtual community usage and development in the future.

5. References

Aberdeen Group (2000) 'Web Analytics: Translating Clicks into Business'. Aberdeen Group, Inc: Massachusetts: USA

Buresh, S. (2003) 'Your Web Traffic and Your Bottom Line'. Web site: http://www.marketingprofs.com/2/buresh.asp, accessed 18 March 2003.

Fletcher, P, Poon, A., Pearce, B., Comber, P (2002), "Practical Web Traffic Analysis: Standards, Privacy,

Techniques, Results", Glasshaus.

Hewson, C., Yule, P., Laurent, D., Vogle, C. (2003). "Internet Research Methods", Sage Publications.

Hine, C. (2000), "Virtual Ethnography", Sage Publications.

Kilpatrick, I. 'Too many hits are bad for your web site!'. Web site: http://www.contractoruk.co.uk/tech-hits.html, accessed 1 May 2002.

Maxmind Ltd. (2004). "The GeoIP Country database", http://www.maxmind.com/

McLaughlin, M., Goldberg, S., Ellison, N., Lucas, J. (1999), "Measuring Internet Audiences: Patrons of an Online Museum", in Jones, S. (ed.), "Doing Internet Research", Sage Publications.

Nonnecke, B. and Preece, J. (2000), "Lurker Demographics: Counting the Silent." Proceedings of CHI'2000, Hague, The Netherlands.

Paccagnella, L. (1997). "Getting the Seat of Your Pants Dirty: Strategies for Ethnographic Research on Virtual Communities", Journal of Computer Mediated Communication 3 (1) June 1997.

Pink, S. (2000), "'Informants' Who Come 'Home'", in Amit, V. (ed.), "Constructing the Field: Ethnographic Fieldwork in the Contemporary World", Routledge.

Preece, J. (2000). "Online Communities: Designing Usability, Supporting Sociability". John Wiley & Sons.

Preece, J., Maloney-Krichmar, D., Abras, C. (2003). "History and Emergence of Online Communities", In B. Wellman (Ed.) "Encyclopaedia of Community". Berkshire Publishing Group, Sage

Rheingold, H. (1993). "The Virtual Community: Homesteading on the Electronic Frontier", Persueus Publishing.

Schmitt, E., Manning, H., Yolanda, P., and Tong, J. (1999), 'Measuring Web Success'. Forrester Research, Inc: Massachusetts: USA.

Scott, J. (2004), "Time, Place and Cybersapce: Locating the Field in an EU Project"

Sheppard, L., Phippen, A., Furnell, S. (2004). "A Practical Evaluation of Web Analytics", to be Published in Internet Research Journal.

Smith (1999), "Invisible Crowds in Cyberspace: Mapping the Social Structure of the Usenet", in Smith & Kollack (eds.), "Communities in Cyberspace", Routledge.

Sterne & Cutler (2000), "E-Metrics: Business Metrics for the New Economy". http://www.emetrics.org/articiles/whitepaper.html

Thomsen, E. (2002), "OLAP Solutions: Building Multidimensional Information Systems, Second Edition", John Wiley & Sons.

Warren, M. & Skerratt, S. (2003), "The 'Virtual Village' – Rural Community Websites and Webmasters as Agents in Rural Development", in Banks, R. (ed.), "Survey & Statistical Computing IV: The Impact of Technology on the Survey Process". Assocaition for Survey Computing.

Wellman, B., Boase, J., & Chen, W. (2002). "The Networked nature of community: Online and offline.", IT & Society, Vol. 1, Issue 1. Summer 2002.

Whitecross 'From Web Logs to Web Loyalty: managing your customers throughout the customer lifecycle'. Web site at: http://www.whitecross.com, accessed June 2002.