In the name of God

E-search Methodes Strategies Book

QUANTITATIV EDATA GATHERING AND ANALYSISON THE NET
Total number of people online by continent (http://www.nua.ie/surveys/how - many- online/index.html) A weaher-type map showing the latency or time delays experienced on the Inter-net from an historic or real-time view at (http://www.mids.org/weather/) Statistics on the so-called digital divide,or the current state of differentiated use of the Net by groups from different socioeconomic backgrounds,gender,and race at (http://www .internetpublicpublicpolicy .com/digdiv.cfm) An updated list of recent surveys completed on Net behavior and activity by the IDS Corporation . This list links to worldwide information sources covering a wide range of issues, although there seems to be an emphasis on e-commerce applications.(http://www .nua.ie/surveys/index.cgi)

From these statistics wecan gather an overview of Net usage, but usually the e-researcher is more interested in activity on particular sites . For this type of data col-lection , we move to a discussion of Web site analytics.
WEB SITE ANALYTICS OR e-METRICS

The explosion of programming and interest in the provision of online services and access to online resources creates a new arena for e-research. Many questions have emerged as a result. Who is using the site? What resources are they utilizing? How long are they spending on each component of a site? What are participants' percep- tions of the value of the site? What suggestions for site improvement do users have? Are usage patterns different between new and experienced site users? There is obviously no single research tool or methodology that provides answers to these and many other important research questions. The traditional means of assess- ing participants' perceptions, suggestions, and concerns through survey or interview research, coupled with evaluations of outcomes has provided answers to some of these questions. However, for other questions, the online environment itself provides a wealth of relevant data. The analysis of this data is a subset of the emerging (and some- what over-hyped) field of study known as data mining. Two Crows Consulting describes data mining as "a combination of machine learning, statistical analysis, mod- eling techniques and database technology. Data mining finds patterns and subtle rela- tionships in data and infers rules that allow the prediction of future results" (Two Crows Corporation, 1999).

Clearly, the research possibilities for such analysis is great. A few of these benefits include the capacity to identify which activities and resources were used most (and least) frequently, the ability to record the length of time partici-pants spend individually and on average using a particular resource, the ability to adapt the activities in response to data gathered on user behaviors, and the capacity to iden- tify individual and group problems when accessing particular pages. The data from Net-based environments are, in one important sense, more accessible to the e- researcher than data from equivalent non-networked environments in that all interac- tivity, postings, and navigation are automatically recorded by the programs that create
the Net-based environments. This section discusses means and ways to mine data from Net-based environments. All activity that takes place on a Web site or in a virtual environment is normally logged or stored by the program owner. However, the difference between the promise of accurate and meaningful information and the reality of what one finds in the thou- sands of lines of raw data produced by a Web server log has inspired a host of Web analysis applications and even more customized solutions. These applications produce a wide variety of individual and summary data, some of which may be useful to the e- researcher. If, for example, one is studying the use of a dedicated educational suite of Web-enabled software, such as First Class, WebCT, or BlackBoard, the program may itself be gathering and presenting data on user activity. This is almost always the easi- est data to access. However, most of the data collected by these programs is designed for educational purposes, rather than research purposes and, hence, may not be opti- mized for e-research. As such, the e-researcher may need to turn to one of the more sophisticated commercial or freeware tools designed to assist researchers in identifying and measuring the activities that users engage in while visiting a site. Unfortunately, most of these tools are focused clearly on the e-commerce market. The tools available, while interesting and of potential value for some types of e-research, may be too focused on analysis of behavior that is directly related to current or future sales prospects. Concurrently the prices of some of these products reflect their commercial orientation.

Aberdeen Consulting coined the term insight-to-effort ratio in regard to Web analyzer software to highlight the amount of effort required of the e-researcher to extract meaningful insights from the behavior of users of the site. Complex analysis information may be very time consuming and require special programming and data extraction skills. At the lowest level, a researcher can use simple text analysis tools to examine and extract data from the Web logs themselves. However these logs are usu- ally overly detailed and not formatted for ease of understanding or analysis. In all like- lihood,

work spent analyzing raw server logs produces a very low insight-to-effort ratio. Alternatively, an e-researcher may choose to lease or buy a high-end Web ana- lyzer package and achieve insights at low effort, though probably at a high price. The e-researcher's task, then, is to select a set of tools that provides a high insight-to-effort ratio without exceeding our often limited research budgets. Beginning e-researchers might wonder just what type of information a Web server routinely captures. The Australian Web service, VSBWEB, provides a real- timeanalysis of a variety of Web sites that it supports. Reviewing the reports at http://vcsweb.com/logs/ provides a glimpse of the types of information available from the logs of a standard UNIX-based Web server and the output formatted using the open-source analysis program Analog (http://www.analog.cx/).The basic features of Web analysis programs relevant to researchers include: Number of hits at specific pages. Amount of time between hits, thus indicating the time visitors spend at each page. . Reviews of the path followed by subjects through particular educational sites.

QUANTITATIVE DATD GATHERING AND ANALYSIS ON THE NET

Demographics provided by a Web log such as type of browser used, domain name location.

The maintenance issues include:

Number of errors of any kind encountered by users
Link verification (should support a wide variety of links-http, ftp, mailto, image, applet, etc.)
Site maps generated in Rich Descriptio Format (RDF)
Orphaned pages-those that are no longer connected to other pages on the Web site A graphical view of individual and summary statistics of navigation through your site
Pages with slow download times
Site reliability-when and for how long the site was not responding to requests for information
A log of search items found in help or directory searches

These machine-gathered data are frequently combined with information provided by the user-typically when they first register at the site or through a standard educational registration process. In educational applications the researcher may have access to other demographic information including grades, prerequisite accomplishments, and scores on pre- and post-tests. Access to

this personal information is of course con- trolled through ethical constraints and the e-researcher must obtain informed consent from participants (see Chapter 5). Ife-research is being conducted on sites where per- sonal information is not being gathered, it is still important to inform users of what information is being gathered and for what purposes. This information should be posted prominently in a privacy policy accessible from the first page a participant is likely toencounter. For help in creating such a policy, or to have your site assessed and credentialed as one that maintains privacy controls, you may wish to contact a non- profit privacy organization, such as www.truste.org. The process of analyzing Web logs can be tedious as the volume and amount of irrelevant data translates into a great deal of preprocessing before analysis can com- mence. Zaiane (2001) lists the major steps in the analysis of educational Web logs:

Remove irrelevant entries. Identify access sessions (to determine individual users). Map access log entries to learning activities. Complete traversal paths (what pages did the user request and in what order). Group access sessions by learner to identify learning sessions. Integrate data with other data about learners and groups of learners. (p. 61)

Fortunately, many applications require participants to log in, so that the activities of different users can be uniquely identified. This login identification is kept on the user's machine and information is passed to the Web server through the appendage of a cookie.