TOWARS WEB SITE USER'S PROFILE
Log File Analysis

Ivo Pierozzi Jr., Eliane Gonçalves Gomes, Maria de Cléofas Faggion Alencar, Carlos Alberto de Carvalho

Satellite Monitoring Research Center, Brazilian Agricultural Research Corporation, Av. Dr. Júlio Soares de Arruda, 803, Parque São Quirino, 13088-300, Campínas, SP, Brazil

Email: ivo@cnpm.embrapa.br, eliane@cnpm.embrapa.br, cleo@cnpm.embrapa.br, calberto@cnpm.embrapa.br

Keywords: Log file analysis, Usage Dynamics, Performance

Abstract: The Internet is a remote, innovative, extremely dynamic and widely accessible communication medium. As in all other human communication formats, we observe the development and adoption of its own language, inherent to its multimedia aspects. The Embrapa Satellite Monitoring is using the Internet as a dissemination medium of its research results and interaction with clients, partners and *web site* users for more than one decade. In order to evaluate the *web site* usage and performance of the e-communication system the Webalizer software has been used to track and to calculate statistics based on web server log file analysis. The objective of the study is to analyze the data and evaluate the indicators related to requests origin (search *string*, country, time), actions performed by users (entry pages, agents) and system performance (error messages). It will help to remodel the *web site* design to improve the interaction dynamics and also develop a customized log file analyser. This tool would retrieve coherent and real information.

1 INTRODUCTION

The Internet, undoubtedly, has become the most innovate, democratic and comprehensive communication medium developed so far. Like all other forms of human interactions, the Internet develops and adopts its own languages and codes, inherent to its multimedia and interactive nature, with the possibilities of real time connections, reuniting in a single media, images, sounds and texts.

Embrapa Satellite Monitoring (Satellite Monitoring Research Center, Brazilian Agricultural Research Corporation – Embrapa) provides consulting to state institutions and private companies in agricultural and cattle raising and environmental issues and challenges, by means of research, adaptation, evaluation and availability of knowledge and information originated from geotechnologies applications. Since 1991, it has been using the Internet not only as a passive media for diffusion of knowledge and of information generated by its research projects (Pierozzi Jr. et al., 2000). Moreover, it has invested and followed the rapid technological pace of the sector, transforming its web site, from a mere shop window of results to a dynamic and active channel of contact with users, clients and partners, who demand products, services and actions of satellite monitoring.

Aware of the importance of opinion, expectation and satisfaction of the public, in relation to content and form of information made available, some initiatives towards evaluation of the website usage were implemented. Web site access monitoring by web trackings – a mechanism for log file analysis – was one of such initiatives. These tools generate several reports for analysis and understanding of variables related to the dynamics of web site usage. Log files offer web sites managers a substantial level of details about the visitors, which presently are rather diversified and qualitatively used (Peters, 1993; Lee and Heller, 1997; Bertot et al., 1997; Yu and Apps, 2000; Murphy et al., 2001; Hochheiser and Shneiderman, 2001).

This paper presents and discusses the results of three and a half years of monitoring of web site usage of Embrapa Satellite Monitoring (ESM), aiming at tracing a profile of its dynamics and obtaining indicators to adjust and improve the design, availability and maintenance of information. To do so, data about the dynamics and origin of usage, usage agency and web site performance were used.

2 METHODS

The data for evaluation of ESM web site usage were those of historical series of reports generated by Webalizer between January 1999 to July 2002. Webalizer (http://www.mrunix.net/webalizer/) is a free software application for analysis of log files, that generates monthly, daily and hourly statistics of usage, ranked according to indicators such as, for instance, hit number, country of origin, search string, and so on. Barrett (1997) proposed the definitions for such indications. Further definitions can be found in Bacalla (1997).

When the decision was taken about using a tool for log file analysis, the Webalizer application was already active in ESM. To keep the objectivity, homogeneity and coherence of information, this tool was not replaced. Due to its peculiarities that limit the range of possible conclusions about the web site usage, specific details can only be obtained through the direct analysis of log files.

Among the set of indicators made available by Webalizer, only were used those that allowed us to evaluate aspects concerning the origin of access, actions executed by users and system performance in terms of response to users' requests. These results can enable the redesign of the web site, in order to improve the dynamics of interaction and to promote the development of a proper approach to log analysis.

3 RESULTS AND DISCUSSION

The usage of ESM web site was analyzed according to the following event composition, whose set determined the specific goals of this study:

- dynamics and origins of usage: events related to the temporal dimension, expressed in terms of monthly, daily and hourly accesses; events related to search string and to the country of origin of the access.
- usage agency: represented by entry pages of higher access rate and most used agents.
- web site performance: inferred by status code (hits and access errors).

3.1 Usage Dynamics and Origins

3.1.1 Temporal analysis

By analyzing the general dynamics of ESM web site usage during the period under consideration, it can be observed that in the first two and a half years the number of hits remained in a rather constant plateau, around 100.000 pages accessed by month, even though a slight yearly growth trend could be seen.

It was observed that the number of hits changed from around a monthly average of 71.732 pages accessed in 1999, to a number of 586.296 in November 2001, corresponding to a different pattern of growth observed so far. Around this time, the first information about the Ecological and Economic Zoning of the State of Maranhão and the release of the CD ROM's collection "Brazil seen from space", two works with a major national repercussion and of general interest of society. From this period on, the trend noticed was a rather significant growth of hits, corroborating the correctness of the institutional strategy adopted regarding the diffusion in the Net of its research results.

July 2002 witnessed a peak in the number of hit pages, resulting from the launching of the CD ROM collection "Brazil seen from space" and the availability of this theme in the web site www.cdbrasil.cnpm.embrapa.br. The subject has awaken an enormous interest in the public in general as the collection is a pioneer work in Brazil: satellite images were reunited and treated digitally in the shape of mosaic to allow the product to represent the surface of the earth as seen from space.

Other aspects observed in the temporal pattern of usage is that, clearly, in the months from December to February, and in the month of July, the access drop to nearly half. These months correspond to the major school vacation period in Brazil. This concentration of use in the months of higher school activity and of the productive sector in general, seem to indicate that the user of the web site concentrates more specifically in the search of information related to the profile of the agency of Embrapa Satellite Monitoring and related issues.

The analysis of daily usage dynamics supports this idea. If can be seen clearly that hits concentrate in working days and working hours, with a clear-cut reduction during lunch time, evenings and the weekends.

3.1.2 Search strings

Web users' visits to the ESM web site can occur in a straightforward manner, if the www.cnpm.embrapa.br URL is used or in an indirect way, if they use strings or key-words, launched in search mechanisms, or by links from other sites directed towards ESM.

The analysis of strings search registered by Webalizer show that the search mechanisms should deal well with users' requests to indicate ESM as reference to web sites about vegetation fires (www.queimadas.cnpm.embrapa.br), given that this

theme has aroused a significant attention national and internationally.

Indications about search mechanism coherence are also revealed when they point to the web site, answering requests to strings such as "embrapa" and "satélite" [satellite]. However, in relation to the string "morcego" [bat] its significant occurrence is explained by the presence in the web site of environmental education home pages. These pages discuss the environmental impact of farming on the wildlife. This specific content seems to be fairly accessed by a public interested in information and pictures about wild animals.

This analysis can help web designers to use strategically the content of pages, choosing words that are more relevant as indicators to be indexed.

3.1.3 Country of origin and access hits

Qualitatively, users from more than 173 countries have accessed the site. From a quantitative point of view, there is a clear pattern, indicating that the origin of access hits is mainly from Brazilian users: more than 60% of the total number of access hits in the period under study. Next, there a significant amount of access whose origin could not be determined and it remained stable between 20 to 30% the total number of hits.

The United States are responsible for the higher rate of hits for a foreign country, occupying the third rank, although they have never reached more than 10% of the total number of access hits. All the rest, encompassing the other countries identified, have never amounted to 5% of that indicator.

These data show a strong national presence of the web site, pointing to a necessary care and attention, by its designers and managers in terms of adequacy of Portuguese and a low priority to dissemination of information in other idioms, at least in a short or medium term.

3.2 Usage agency

Important information to the management of web site content can be obtained by monitoring users' actions executed from the moment they access it. The path undertaken and the files recovered remotely can provide clues or even ratify the level of interest aroused in the public by a given set of information.

However, statistics organized by Webalizer allow a limited analysis of these actions. Among the situations that could be treated, we present, next, data referring to web site entry home pages and agents used in the visits.

3.2.1 Entry home pages

The analysis of information about entry home pages demonstrates that users explore the content of the web site in a variety of ways. The home page of Embrapa Satellite Monitoring (http://www.cnpm.embrapa.br/) and subjects about orbital monitoring for vegetation fires in Brazil (http://www.queimadas.cnpm.embrapa.br/) display pages that are continuously accessed during all months of the years. Other subjects, such as the study of locusts in Mato Grosso (http://www.cnpm.embrapa.br/projetos/gafa_mt/index.html), works in the municipalities of Jaguariúna, state of São Paulo (http://www.cnpm.embrapa.br/projetos/jaguar/index.html) and of Machadinho d'Oeste, state of Rondônia (http://www.cnpm.embrapa.br/projetos/machadinho/index.html), home pages with information about satellite used in agroecological research (http://www.cnpm.embrapa.br/vp/saibamais/index.html) present, from year to year, a random variation in the rates of access, reflecting access by users just interested momentarily in those information.

It can be also observed themes that clearly generate impact on public interest, since, as soon as they are made available, they become systematically accessed, such as, for instance, work carried out in the Demene River, in the border of the states of Amazon and Roraima (http://www.cnpm.embrapa.br/projetos/demene/index.html) and "Brazil seen from space" CD ROMs (http://www.cdbrasil.cnpm.embrapa.br/).

3.2.2 User agents

Agents were grouped in 4 categories, namely: undefined, browser (Netscape, Internet Explorer, etc.), robots (crawlers, spiders, link checkers, proxys) and mirroring (they capture the site to enable off-line navigation).

The analysis of results shows that almost all users of our web use browser agents. These results seem to indicate that the common user is a member of academia (student, professor or researcher), staff of government agencies or of NGOs, entrepreneurs, and members of organized civil society. These users, most likely, employ those browsers that are well known and disseminated in the market due the easy of acquisition, setup and familiarity of usage.

In the case of robot agents, an increase in their participation in the total number of access hits can be observed for the year 2000, what could be explained by a large amount of information made available specially during that year, such as for instance, information about the National Campaign on Alternative Practices to Slash and Burn

Agriculture and the collection "Brazil seen from space".

Most interesting, however, is the significant and steady increase of the share of mirroring agent since 2001. This fact is due to data of the project "Brazil seen from space" made available, what has aroused public interest in obtaining information, stimulating download of the whole set of files to their personal computers.

3.3 Web site performance

The temporal evolution of web site performance was evaluated according to status codes. The most relevant point, in relation to performance of ESM web site, is the steady percentage of successful access codes during the period under study. It is practically always near 100%. At the same time, there are no registers for 500 error-type message. The unit high investments and constant concerns about update and maintenance of infra-structure and the availability of operational resources have warranted the availability of external access to information round the clock, seven day a week.

The increase in the year 2000 of error-type 400 [syntax errors] can be explained by the change underwent in the name of the institution and, consequently, of the web site home pages URLs. In 2000, Embrapa Satellite Monitoring institutional name was transformed from "Núcleo de Monitoramento Ambiental e de Recursos Naturais por Satélite" (www.nma.embrapa.br), to "Centro Nacional de Pesquisa de Monitoramento por Satélite" (www.cnpm.embrapa.br). This change might have caused a considerable increase in the type-error 400 when users, accustomed to accessing the web sites with syntaxes containing the string "nma", did not succeed in logging in.

4 CONCLUSIONS

The analysis of data gathered and organized by Webalizer, allow drawing some trends and patterns in the usage of information made available, via Web, by ESM. Some indication about the system functionality and performance could also be inferred.

This analysis open up a rich perspective to explore and characterize user's profile, demanding the development of customized procedures to log file analysis according to predetermined interested, thus consolidating itself as one of the steps in the process of conception, availability and usage of information.

It is important to notice that the use of data mining, attributes inspection and others testing techniques could be used to support the analysis of log files and to help to remodel the web site design.

Consequently, some ideas are advanced in an exploratory scenario:

- to separate web site sections aiming at determining access distribution;
- to intercept the most accessed entries and compare them with the log;
- to promote structural change based on indexing rules in the mechanisms of search and robots;
- to identify, by occurrence of *mirroring* agents, opportunities to download complete sections;
- to develop our own approach to web server log file analysis;
- to adjust the presentation of access statistics to different public.

REFERENCES

Baccala, B., 1997. Connected: An Internet Encyclopedia. Available at: <http://www.freesoft.org/CIE/index.htm>. Accessed in: 25 Sept. 2002.

Barret, B.L., 1997. The Webalizer - A web server log file analysis tool. 1997. Available at: <ftp://ftp.mrunix.net/pub/webalizer/README>. Accessed in: 15 July 2002.

Bertot, J. C.; McClure, C. R.; Moen, W. E.; Rubbin, J., 1997. Web usage statistics: measurement issues and analytical techniques. *Government Information Quarterly*, v. 14, n. 4, p. 375-395.

Lee, S.; Heller, R. S., 1997. Use of keystroke log file to evaluate an interactive computer system in a museum setting. *Computers Education*, v. 29, n. 2/3, p. 89-101.

Murphy, J.; Hofacker, C. H.; Bennett, M., 2001. Website-generated market-research data: tracing the tracks behind visitors. *Cornell Hotel and Restaurant Administration Quarterly*, p. 82-91, February.

Peters, T. A., 1993. The history and development of transaction log analysis. *Library Hi Tech*, v. 11. n. 2, p. 41-50.

Pierozzi Jr., I.; Caputi, E.; Filardi, A. L., 2000. A Internet como veículo de comunicação, difusão de resultados e imagem institucional na Embrapa Monitoramento por Satélite: I. Infra-estrutura e funcionamento. Campinas: Embrapa Monitoramento por Satélite, 25p.

Yu, L.; Apps, A., 2000. Studying e-journal user behavior using log files: the experience of SuperJournal. *Library & Information Science Research*, v. 22, n. 3, p. 311-338.