Informetrics and the World Wide Web: a case study and discussion

J. Stephen Downie
Graduate School of Library and Information Science
University of Western Ontario
jdownie@julian.uwo

This paper discusses how the application of informetric modelling techniques and principles offers a powerful set of analytic tools for empirically grounding one's understanding of World Wide Web interactions. Data collected from the transmission statistics of a non-profit Web site are presented to illustrate the usefulness of informetric analyses for both scholars and practitioners. A discussion of possible analytic problems and pitfalls concludes the paper.

Introduction

How does one know whether a particular World Wide Web (WWW) site is effective? One possible method of answering this question is to perform a range of informetric analyses.[1] However, before one can perform any informetric analysis, one has to know what to measure. One must understand the theoretical implications and the limitations of the things chosen to be measured. This paper introduces the reader to the types of measurements, and thus the informetric analyses, made possible through standard WWW server logging procedures. Special attention is drawn toward warning the reader about potentially misleading data and other analytic pitfalls.

Case study: INFACT Canada's Web site statistics

To illustrate this paper usage statistics from the INFACT Canada Web site are used. These statistics summarize data from 15 April to 22 April 1996. INFACT Canada, Inc. is a non-profit, nongovernmental organization that advocates for the breastfeeding of infants. INFACT Canada's Web site is located at http://www.io.org/~infacto and has been running since June of 1995. The site contains hypertext versions of INFACT Canada's recent monthly newsletters, a collection of breastfeeding abstracts culled from the medical literature, a collection of award-winning advocacy posters, and many individual articles written by the INFACT Canada staff. Also, a variety of Common Gateway Interface (CGI) forms allow users to contact INFACT Canada, order support material, and perform keyword searches. Documents are served from IO

Canada (http://www.io.org), a Toronto-based Internet service provider (ISP), which uses the NCSA HTTPD 1.4 WWW server. The author and C. Diane Smith are the current INFACT Canada Webmasters.

The NCSA HTTPD 1.4 server provides the ability to create a series of log files[2]—the ErrorLog, TransferLog, AgentLog, and RefererLog. The ErrorLog is used to record information about such errors as requests for missing or broken files, attempts to access locked directories, and erratic CGI results. The TransferLog stores information on all the files requested including their names and sizes, the time of the requests, and the location of the requesters using Internet Protocol (IP) addressing. The AgentLog contains information on all the client browsers that have accessed the site. The RefererLog keeps track of the Uniform Resource Locators (URL) from which the requester has connected to a particular page.

IO Canada allows individual users indirect access to its TransferLog through an intermediary program called "wwwstat 1.0" (Fielding), which summarizes the information in the TransferLog, thus giving a Webmaster daily and hourly transmission statistics, total transfers by client domain and reversed subdomain, and total transfers of each file. At IO, wwwstat 1.0 summarizes the data into weekly usage reports. Table 1 presents a representative, albeit greatly abbreviated, sampling of the wwwstat 1.0 output tables.

The most succinct of the wwwstat 1.0 tables is the weekly summary table reproduced in Table 2. At first glance it appears that the INFACT Canada Web site has had a productive week. The 41,895,874 bytes of information transferred is an encouraging sign. IO Canada charges a "heavy-usage fee" for any site that moves more than 50,000,000 bytes per week, so by this standard the INFACT Canada site is entering the big leagues. It turns out, however, that these data are misleading and give only a partial indication of the true success of the site. As the paper progresses, the extent of the distortion will become apparent.

Possible analyses

Before investigators or Webmasters start wading through the mountains of data that the standard log files can contain, they should have some idea of what it is they want to know. To this end, this section presents a brief taxonomy of potentially useful analyses. Analyses can be organized into three general groupings: user-based, file-based, and session-based. The shortcomings and pitfalls of these analyses are discussed in the next section.

User-based analyses

While information about the nature of individual users is not readily discernible

Table 1. Sample extracts from the wwwwstat 1.0 output

Total transfers by client domain (heavily abridged)

%Reqs	%Byte	Bytes sent	Requests	Domain	Country
0.29	0.53	222010	11	ae	U.A.E/
0.21	0.02	6447	8	ar	Argentina
0.08	0.00	294	3	at	Austria
1.50	0.77	323424	56	au	Australia
4.25	3.32	1391319	159	ca	Canada
26.73	29.56	12383775	1000	com	Commercial
1.18	0.72	302565	44	org	Non-profit
18.93	19.35	8108463	708	unresolved	????

Total transfers by archived section (heavily abridged)

%Reqs	%Byte	Bytes sent	Requests	File
0.21	0.05	22168	8	/~infacto/
0.27	0.07	27710	10	/~infacto/Overview.html
0.03	0.01	4527	1	/~infacto/abs.htm
10.91	3.12	1305984	408	/~infacto/smlogo.gif
0.40	0.21	86610	15	/~infacto/keywords.htm
0.05	0.01	2348	2	/~infacto/wwwais.cgi
0.03	0.00	341	1	Error 302 Redirected
1.15	0.03	12068	43	Error 404 Not Found

Total transfers by reversed subdomain (heavily abridged).

%Bytes	%Requests	Bytes sent	Requests	Reversed subdomain
18.93	19.39	8108463	708	Unresolved
0.29	0.53	222010	11	ae.net.emirates
0.05	0.01	6245	2	ca.brocku.cosc
0.37	0.66	277842	14	ca.carleton.ccs
0.05	0.18	76618	2	ca.uwo.slip
0.91	0,30	126358	34	com.inktomi
4.46	5.48	2445708	167	com.aol.proxy

Table 2. Totals for summary period: 15 April 1996 to 22 April 1996

Files Transmitted	3741
Bytes Transmitted	41895874
Average Files Transmitted Daily	468
Average Bytes Transmitted Daily	5236984

from log files, it is still possible to gain some understanding of the users as an aggregate. Below are several types of useful analyses.

Use by country

One of the mandates of INFACT Canada is to disseminate accurate breastfeeding information to as many countries as possible. During the week under discussion, 31 nations had users access the site. As one would expect, the developed world (United States, Canada, etc.) is the most heavily represented; however, such lesser-developed nations as Malaysia, Brazil, Thailand, and Mexico are also represented. An analysis that verified a wide international dissemination of information would be useful for INFACT Canada in garnering international development grants.

Use by organization type

If one knows something of the types of organizations through which access is made, then one can begin to understand the nature of the users of the information that the site provides. Of course, this is an analysis fraught with assumptions, but nevertheless there are times when one wishes to have at least a cursory understanding of the organizational affiliations of the users. For users from the United States this is relatively easy to achieve. The domain suffixes in the United States simplify identification (e.g., .edu = educational institution; .com = commercial organization; .gov = federal government institution). For users from the rest of the world it is a little more difficult but not impossible. By submitting each IP address in question to the appropriate X.500 or WHOIS directory, one can discern the type of organization associated with the address. For example, the best place to search for information on Canadian domain names is "gopher://nstn.gopher.ca7006/7n".

Use by browser type

In an age where WWW standards are anything but standard, analysis of the AgentLog is highly recommended. In November 1994 Netscape browsers accounted for only 20% of WWW accesses, but by January 1995 Netscape's share had increased to 80% (Magid et al. 1995, 14). Netscape is noted for pushing the standards envelope through its introduction of such innovations as the <IMG>, <CENTER>, and <BLINK> tags. By monitoring the browser stats one can get feeling for when to introduce new non-standard Hypertext Mark-Up Language (HTML) extensions such as frames, server-pushes or JAVA scripts.

Cooper's "odds-payoff rule of indexing" (Cooper 1978) could be adapted to aid the Webmaster in deciding what new features to implement or old ones to discontinue. Losee (1990, 163) expresses Cooper's odds-payoff rule as:

$$\frac{n}{p} < \left|\frac{E(U+)}{E(U-)}\right|$$

where n is the number of users experiencing negative utilities and p is the number of users experiencing positive utilities as caused by the selection of a given item through the use of a given indexing term. $E(U–)$ is the expected utility for the n users experiencing negative utilities and $E(U+)$ is the expected utility for the p users experiencing positive utilities. If the ratio of (n/p) is less than the absolute value of [$E(U+)/E(U–)$]) then a given item should be indexed by the term under consideration. By substituting "feature" for "indexing term" Cooper's equation becomes, when combined with the browser data, a powerful tool in deciding whether to implement a new HTML feature within a Web site.

File-based analyses

The principal components that make up a typical Web site are text files, image files, and CGI script files. Thoughtful analyses of the information found in the TransferLog can help a Webmaster better manage the Web site. Low usage numbers for a particular file could indicate a poorly indexed document, a document that is located too many levels removed from the home page, or an erratic CGI script. Special attention should be paid to the error codes recorded. For example, Code 404 errors (see Figure 1) are a useful indication that links within the Web site contain malformed URLs. Whatever aspect of the TransferLog one wishes to analyse, one should decide whether a byte-based or request-based analysis is appropriate.

Byte-based analyses

The strongest reason for performing a byte-based analysis is that it gives the best indication of the actual work being done by the server. A byte-based analysis of the INFACT Canada data brought to the Webmasters' attention that image files constituted 46.53% (1,740) of the requested files but made up a remarkable 84.04% of the bytes transferred (35,214,748 bytes). Similarly, text files constituted 47.96% (1789) of the requested files but made up only 14.79% of the bytes transferred (6,202,194 bytes). If server bandwidth is at a premium, then modifying the size of, or access to, troublesome image files can pay off handsomely.

Request-based analyses

To gain an understanding of what users are (or are not) actually looking at, request-based analyses are invaluable. At the INFACT Canada Web site the image files are, for the most part, secondary to the text. The images of the award-winning posters that are offered for sale to the public are a probable exception this assertion. Notwithstanding the poster images INFACT Canada spends a substantial amount of money researching, writing, and then mounting its articles. If an article is receiving zero or few requests, it is important for the Webmasters to determine that it is not a Web site structural problem that is limiting access. Broken URLs, improper referencing, and poor interface design can all be flagged through a request-based analysis.

A rank-frequency analysis of request data is helpful for modelling use at a Web site (Table 3). To perform this type of analysis, one orders the request data according to the request counts, denoted as *f(x)*, with the most frequent being ranked *(x)* first. Ties in request counts are settled by assigning the highest rank associated with a given count. For example, if the count of 30 occurs three times at ranks 14, 15, and 16, then assign the rank of 16 to the count of 30. Once this is done, one should notice that the head of the distribution is quite large. The INFACT Canada request data (text files only) has been analysed. The home page is ranked first, with 304 requests. Moving down the ranks, one should notice that the counts start to drop off quite sharply and quickly tail off to a great many items that were requested only once. A cursory investigation of the documents represented in the tail of the distribution revealed that many of the single-request items belong to the collection of medical abstracts. As access to these documents is available only via a CGI search-engine interface, the analysis of the tail suggests that the Webmasters should (1) redesign the search-engine interface or (2) provide greater access through the possible creation of a table of contents page.

Table 3. Rank-frequency data.

Rank	Frequency	Rank	Frequency	Rank	Frequency
1	304	13	35	34	12
2	183	14	31	37	11
3	125	17	30	40	10
4	99	19	25	44	9
5	87	20	20	45	8
6	84	23	19	50	6
7	78	25	18	55	5
8	74	26	17	48	564
9	48	27	16	64	3
10	44	28	15	76	2
11	40	30	14	116	1
12	37	33	13		

The astute reader will note that the distribution described above has a classic Zipfian shape. The rank-frequency data were indeed fitted to a Zipf distribution by means of the form of the Zipf equation that most readily lends itself to linear regression:

$$\ln f(x) = \ln(a) - b\ln(x)$$

where a and b are parameters of the model dependant on the data.

The linear regression determined that $a = 558.96$ and $b = 1.15$. The Zipf model was deemed to fit as the regression was statistically significant (p 0.01) with $N = 35$ and $F = 554.77$.

Session-based analyses

If one is fortunate enough to have access to a server's ReferLog (the author currently does not) a number of informative analyses can be performed. Essentially, session-based analyses are a form of citation analysis. By investigating the originating URL of a request, a Webmaster can determine (1) how users are navigating through the Web site and (2) where the Web site fits into the WWW as a whole.

Search-engine analyses

Most thoughtful Webmasters have submitted their Web site URLs to one or more of the large WWW indexing services such as OpenText (http://www.opentext.com), Lycos (http://www.lycos.com), or Yahoo (http://www.yahoo.com). By analysing the number of referrals made from these search-engines, it is possible to decide whether one's Web site has been effectively indexed. Each indexing service has a different search algorithm and ranking procedure. Some, like Yahoo, allow Webmasters to classify their sites and submit short abstracts. Low numbers of referrals from the major search-engines should inform the Webmaster that it might be necessary to (1) add more topical keywords to important pages; (2) reconsider the classification of the Web site; and/or (3) rewrite descriptive abstracts where applicable. Once the corrective actions have been performed, the URLs should be resubmitted to the indexing services.

Internal search-engine use should also be analysed. Low referrals from internal search-engines can indicate poor interface design, poor indexing, or a even a dislike of search-engines among the Web site's users. Comparing the ratio of referrals from internal pages to the referrals from internal search-engines should aid the Webmaster in determining the effectiveness of internal search-engine design. Application of the aforementioned "odds-payoff rule" might assist the Webmaster in deciding whether to continue internal search-engine capabilities.

Cohort identification

As a Web site becomes established, other Webmasters with related or similar information place links from their sites to the Web site. Referring URLs that appear frequently in the ReferLog should be investigated on the basis of the assumption that the referring URL points back to a site of similar or related information. Such sites should be visited and, if found to be acceptable, their respective Webmasters contacted. By contacting the Webmasters of heavily referring sites, a Webmaster can establish connections to groups that previously might not have been considered cohorts. Thus careful analysis of the ReferLog can become an invaluable collection development tool. Reciprocal links should also be provided back to these referring sites.

Chain analyses

By means of the ReferLog it should be possible to map out a chain of referrals from document to document for each user. For example, imagine a user who starts on the home page (Doc1), jumps to an article (Doc2), jumps again to the search-engine (Doc3), and then returns to the home page (Doc1). If each document is considered a state and each jump a transition, then the system could be modelled as a Markov process (see Egghe and Rousseau 1990, 175–82, for an in-depth description). Watters defines the Markov model as:

> ...a probabilistic model in which the probabilities of future events depend only on the current state and not on how the model got to that state. This means that different systems reaching a state by different means will have the same probabilities with respect to future events. (1992, 134)

Markov processes have been used to describe library circulation by Morse (1968), Chen (1976), and Beheshti and Tague (1984). If Web site usage is shown to be a Markov process, then it would be possible to predict (probabilistically) the paths taken by users. Investigating the possible causes of the transition probabilities should aid the Webmaster in restructuring the Web site to increase efficiency.

Investigator beware: dirty data and other pitfalls

Many things stand between an investigator and meaningful analyses of Web site data. This section is intended to forewarn the investigator of the types problems that might be encountered. Problems can be organized into three general classes: data distortion, user identification, and policy issues.

Data distortion

Although computer log files do not lie, they certainly do not tell the whole truth. The greatest challenge facing the analyst is making sure that the numbers mean what the analyst thinks they mean, especially summary data. There is no substitute for going over the data by hand in an attempt to determine the presence of data-distorting influences. Once identified, the analyst is left to decide how to counteract the distortions.

Spider infestation

Many of the WWW indexing services create their indexes by operating special programs called "spiders" that reach out to Web sites and then explore each link that they encounter. The spiders then submit the pages that they find back to the indexing machine. The INFACT Canada Web site was visited during the summary period by the Inktomi indexing spider (http://www.inktomi.com). This spider requested a total of 34 files (126,358 bytes). One defence against having the spiders distort the data is to maintain a list of spider URLs. Such a list could be use to filter out (or at least flag) data associated with those URLs.

Maintenance issues

One problem with log files is that they log just about everything, including accesses made by the Webmaster during the course of maintaining the site. When a Webmaster fixes a certain document or checks the search-engine, the request numbers for those items tend to skyrocket. Since most analyses are conducted with an eye toward determining user (not Webmaster) patterns, the Webmaster should be meticulous in maintaining a personal log (probably paper-based) of accesses made to the Web site. Also, if a time-based study of Web site use is undertaken, the investigator should be aware of possible maintenance shutdowns of the Web site's server.

Imbedded images

The greatest single distorter of both the request-count and the bytes-transferred data is the presence of imbedded images. When a user requests a given document the server finds the document, serves it to the requester, and then records the request. If that document contains an imbedded image, the server opens up another session to serve the image and records the request. If the document contains multiple images, then multiple requests are recorded (one for each image).

The INFACT Canada Web site statistics are distorted precisely because of the images imbedded within its pages. There is an imbedded image file at the site called "smlogo.gif". This file functions as a kind of logo that sits in the upper-right corner of almost every page. It could be successfully argued that no user ever "requested" this file, as it is simply attached to each document. However, the analyst must deal with the fact that this one file represents 19.23% of requests (719) and 5.49% of bytes transferred (2,298,622 bytes).

User identification

When conducting analyses based on identifying users, whether by nationality, by institution, or as individuals, many problems can arise. A substantial number of IP addresses exist that the WWW server can not decipher. These are recorded as "Unresolved" in the summary statistics. Using the X.500 or WHOIS directories to ascertain the type of institution associated with a given IP address is only effective if the institution has properly registered its IP address with the appropriate database. For example, the author's searches for information on certain IP addresses originating in Brazil have yet to be successful. To further complicate matters, large ISPs such as AOL (http://www.aol.com) or Compuserve (http://www.compuserve.com) have so many different users that attempting to classifying their users as a group is fruitless. For example, where exactly is an AOL user, given that AOL offers Internet access from across North America?

The use of IP address pools for Serial Line Internet Protocol (SLIP) and Point to Point Protocol (PPP) makes it difficult to associate an individual with a given IP address. For example, the author has logged into the INFACT Canada site as ts17-7.slip.uwo.ca, lost the telephone connection, and then immediately regained access as ts17-4.slip.uwo.ca. Also, some early versions of web browsers are known to submit faulty information (Magid et al. 1995, 80). Simply put, accurate data for user-based analyses are very difficult to obtain, and great care must be taken before results based on such analyses yield useful information.

Policy issues

Before any analysis can take place, the data have to exist and the analyst must have access to them. This is not as simple as it sounds. The needs of Webmasters and those of server administrators are not the same. While the thoughtful Webmaster would like access to as many log files as possible, the server administrator has maintenance and security issues with which to grapple.

Some administrators deliberately do not create all of the log files that the various WWW servers can generate. Log files for large WWW servers can become enormous. Not only do they increase administrative overhead by taking up valuable processing time and disk space, they also have to be maintained. Giving individual users access to log files produces even more processor overhead and security problems. For example, IO Canada does not (to the author's knowledge) use the ReferLog or AgentLog capabilities of the NCSA HTTPD 1.4 server.

There are three possible solutions to the problem of log access. First, the analyst could persuade the server administrator of the usefulness of the desired analyses and convince said administrator to begin logging the necessary data. Second, if CGI scripting is allowed on the server, then each document could be created as a CGI script that logs the information submitted by the requester's browser before serving the document. (While possible, this is the least efficient solution, as CGI scripting is trouble-ridden and increases processor overhead). Third, and probably the most efficient solution (from a data collection viewpoint), the Web site could be remounted on a machine where the Webmaster is also the server administrator.

Summary and suggestions

This paper has introduced the reader to potential informetric analyses of Web site log data and the uses to which such analyses can be put. Attention has been brought to bear on the possible distorting factors that might hamper analysis or influence interpretation of the data.

Should the reader be interested in learning more about this subject, the author highly recommends the following two WWW sources:

- http://www.piperinfo.com/piper/9512/usage.html
  This site contains an excellent article on WWW log analysis. Useful for pointers to other sites that explore analysis subtopics in greater detail.
- http://union.ncsa.uiuc.edu/HyperNews/get/www/log-analyzers.html
  An excellent source for locating programs that are tailor-made to aid in the analysis of each type of log file.

Notes

[1] The author uses "informetrics" as defined by Egghe and Rousseau. Informetrics is "...the measurement, hence also the mathematical theory and modelling of all aspects of information and the storage and retrieval of information" (Egghe and Rousseau 1990, 1).

[2] Discussion is limited to the logging capabilities of the NCSA HTTPD 1.4 server, as 90% of sites are served by NCSA servers (Magid et al. 1995, 63).

References

Beheshti, J., and J. Tague. 1984. Morse's Markov model of book use revisited. *Journal of the American Society for Information Science* 35: 259–67.

Cooper, W.S. 1978. Indexing documents by gedanken experiment. *Journal of the American Society for Information Science* 29(30): 395–411.

Egghe, L., and R. Rousseau. 1990. *Introduction to informetrics: quantitative methods in library, documentation and information science.* New York: Elsevier Science Publishers.

Fielding, Roy. n.d. wwwstat 1.0 Available at: http://www.ics.uci.edu/Websoft/wwwstat/.

Losee, Robert M. 1990. *The science of information.* San Diego: Academic Press.

Magid, Jonathan, R. Douglas Matthews, and Paul Jones. 1995. *The web server book.* Chapel Hill: Ventana Press.

Morse, P.M. 1968. *Library effectiveness: a systems approach.* Cambridge, MA: MIT University Press.

Watters, Carolyn. 1992. *Dictionary of information science and technology.* San Diego: Academic Press.