https://doi.org/10.48009/2_iis_2007_7-13

ANALYSIS OF WEB PAGES AND METRICS RELATED TO GLOBAL ENVIRONMENTAL MANAGEMENT

G. Kent Webb, San Jose State University, webb_k@cob.sjsu.edu

ABSTRACT

Much of the media's attention has recently been focused on problems with the global environment and its management. This paper examines related internet sites and focuses on how web metrics can provide information on the priorities and interests of internet users. The conclusions from the data analysis are that keyword search statistics provide good information on user priorities that may be used in place of more expensive survey based research. Keyword search data significantly reproduced the results of a survey used to identify investment banking opportunities in the area of global environmental management. Also, contrary to expectations, the rank order of Google search results were positively correlated with traffic: the higher the traffic, the farther down in the search result.

Keywords: Website Traffic, Business Intelligence Search Engine, Globalization

INTRODUCTION

Climate change, oil dependency, hunger, and epidemics are among the top global environment issues according to a recent survey [15] which also reveals that venture capital firms have responded to these concerns by significantly increasing investments in clean technology, renewable energy sources, and other related technologies. The top opportunities for these venture capital firms as identified by this survey are reported in Table 1.

Table 1: Investment Survey of Top Commercial Applications for Global Environmental Management

Global Environment Issue	Rank
Global Warming	1
Oil Dependency	2
Hunger and malnutrition	3
Dirty Air	4
Dirty Water	5
Over Fishing	6
Epidemics	7
Drug-resistant infections	8
Waste Disposal	9

The economic scale that allows for globalization has also created significant global environmental problems. One product of scale economies, the internet, developed to allow scientists and researchers to share information across wide distances has more recently expanded into a global mass medium. Freely available tools for language translation make information on the internet more accessible. All of the world's media provide a market of ideas and information that can be part of the solution as well as a measure of users' interests and priorities; but the internet is unique in also providing a rich, readily available source of data that can be used for many kinds of analysis.

The growing literature related to measuring the information needs of internet users is often referred to as web metrics [3]. The idea that these measures can provide some insight into future trends is an extension of research suggesting that simple markets can be used to aggregate disperses information into efficient forecasts [16]. Instead of voting with their feet or dollars as in economics markets, internet users can be considered to be voting with their web page visits or searches. Given the readily available information on web page traffic and search statistics, it seems that it might be feasible to count these "votes" and make interferences similar to the survey results summarized in Table 1. The internet data would come from a huge sample but at a much lower cost than for traditional survey results.

Although the term "web metrics" is drawn from the computer science literature and so has been focused on how to deliver the most useful information in a database search, there is also significant interest in the business literature related to what attracts users to web pages. Much of this literature draws data from surveys or log files from individual servers [2, 7, 8, 10].

There have been numerous innovative uses of the internet related to global environmental management. For example, a new page at www.tolweb.org has been designed to coordinate a global effort to maximize local effectiveness in expanding biodiversity. Acclaimed biologist Edwin O. Wilson has identified 10 areas on 2.5 percent of the world's land where 50% of the biodiversity is supported. The

goal here is to coordinate professional and amateurs to optimize management of biodiversity. This page, however, turns up hundreds of pages down in a Google search on the topic. Even when one finds the home page of tolweb.org, it is hard to find the specific information. It is a great idea, but hard to find. One solution to this problem is to create information portals, such as the one underway at www.eol.org, which proposes to create a portal linking all internet information related to species.

Eyob [5] discusses the problems associated with finding information given the explosion of web pages and conducts a regression analysis of the impact of performance, consistency, and optimization on customer satisfaction, suggesting that future research could incorporate more explanatory variables. Shi [14] studies the search engine visibility of sites and concludes that organizations need to consider the type of data that can be collected to "optimize their search engine marketing strategies." Lo and Sedhain [9] find that web page rankings are very sensitive to the "method" of ranking and so suggest using multiple search results and that users go beyond the "30th entry."

This paper considers the use of keyword search statistics as an explanatory variable that can be used for web page design and content strategies as well as for other more general applications such as investment decisions. The paper also examines how reliably search page results relate to users interest as measured by actual web page traffic.

SOURCES OF DATA

The survey of investment opportunities for managing the global environment, summarized in Table 1, was used as a baseline for this research. Other data sources include Google search listings for web pages related to global environmental management, TrafficEstimate.com for home page visit estimates. (this data is also estimated by MetricsMarket.com), and WebFooted.net which provides estimates of keyword searches, broken down by search engine. Individual web pages discovered from the search were also analyzed for content.

Search engines such as Google use a combination of relevance and quality metrics in ranking the responses to user queries. Algorithms to determine relevance depend on keyword term frequencies [17]. The quality of a web page is thought to be related to the hyperlink structure of the page. Brin and Page define the "PageRank" as the probability that a random surfer visits the page [1].

A Google search on "global environmental management" was conducted during March of 2007. The first 350 page URLs that came up from the search were saved for use in this analysis. Monthly traffic was estimated for each page, but one limitation to the analysis was that only traffic to home pages (the home page server) could be estimated. For example, the Wikipedia home page has a huge number of monthly traffic visits, but once the traffic gets inside the home page server, the estimation tools that are available cannot further identify the specific destination of the traffic.

RESULTS

As reported in Table 2, none of the ten most visited sites ranked in the Google top ten. Many of the pages appearing in the Google search results, however, were sub pages on a web site and so their specific traffic could not be estimated. For example, although the web page on global environmental issues at the World Health Organization rated high in the Google search, it is contained in a sub directory related to the environment, not the home page. As a result, the traffic to this specific sub page could not be estimated.

Even with the limitations of the traffic estimates, a clear pattern appears to emerge from the data. For example, the first page listed from the Google page search, www.gefweb.org, received only about 17,000 visits per month. While this organization is a laudable international group providing funding to deal with environmental issues, an informal inspection of the page gives no readily apparent clues about why it is at the top of the search when equally relevant pages getting 100 to 200 times the monthly traffic are buried much farther down in the search results.

The top two pages in terms of traffic, the widely used web site supported by NOAA (National Oceanic and Atmospheric Administration) and the popular educational website at teachearth.com, were far down in the Google search listing, more than 300 pages down the list. Several of the most visited web sites offered information and access to databases. The highest traffic site uncovered , www.nesdis.noaa.gov, "provides timely access to global environmental data from satellites and other sources to promote, protect, & enhance the Nation's economy, security, environment, & quality of life [12]." Other web sites devoted to sharing data include the United Nations site na.unep.org,

Table 2: Top 10 Web Sites by Monthly Traffic and Their Google Search Rank

Traffic Rank	URL	Monthly Traffic	Google Search Rank
1	www.nesdis.noaa.gov	3,968,000	319
2	teachearth.com	2,990,000	317
3	chge.med.harvard.edu/	2,740,500	15
4	sustainabilityscience.org	2,724,400	306
5	www.erb.umich.edu/	1,994,100	214
6	www.sage.wisc.edu/	1,657,700	24
7	na.unep.net/	1,449,200	58
8	environment.newscientist.com/	1,291,600	28
9	www.ceage.vt.edu/	898,400	87
10	nigec.ucdavis.edu/	851,100	20
Based on a sample of 350 pages from Google, 49 for which traffic could be estimated.			

Half of the top ten most visited sites were .edu, related to university research centers on environmental issues. Table 3 gives a breakdown of the URL suffix for the top 350 web pages discovered in the Google search. About 13 percent of the sites were sub pages contained in web sites already identified in the search. For example, a number of different pages for the United Nations sites appeared in the search results.

Although only 18.5 percent of the pages contained a specific suffix that identified it as a country outside of the United States, many of the sites in the .org, edu, and .com were also international sites. The United Nations is a good example, listed as an .org.

Table 3: Summary Statistics for Web Pages Discovered from the Google Search (top 350)

Percentage of Pages from Duplicated Sites in the Search	13.1%
Percentage of Pages with the Suffix	
.org	30.5%
International	18.5%
.com	18.5%
.edu	15.4%
.gov	3.8%
.net	3.5%
.info	0.6%

The keyword search estimate from webfooted.net is designed to help web page designers include the best terminology to attract users of the main search engines looking for information. An analysis of the term "global environment" produced the results reported in Table 4. It turns out that very few, only about 20 per week, searches were done on the larger phrase "global environmental management" compared to the 641 weekly searches estimated for the phrase "global environment". Priorities based on this search were: global health, sustainability, global warming, marine and forest issues.

Table 4: Top Keyword Search Phrases Related to the Global Environment (Weekly Searches)

Keyword or Phrase	Estimated Weekly Searches
Global Environment	641
Environment Global Health Our Perspective	174
Environment Global Sustainable	65
Environment Global Warming	38
Biosphere Environment Global Our Protecting	33
Environment Global Marine	29
Energy Environment Global Resource Science	
Environment Fund Global	28
Environment Forest Global	26
Change Environment Global Science	24
Effects Environment Global Warming	21
Canada Environment Global Warming	19
Non-ecological phrases such as the business environment were excluded from the list. Estimates from www.webfooted.net	

As reported in Table 5, the names of the global environmental issues used in the investment survey are not exactly the phrases most likely to be used in an internet search. For example, few people use the phrase "oil dependency" in their searches, the phrase used in the investment survey from Table 1. Related terms such as solar power with estimated monthly traffic of 5,801, score much higher. The phrase "dirty air filter" scored about 40 percent more traffic that "dirty air". The related phrase users seem to be searching on most commonly is "air pollution" with an estimate of 4,372 per week.

The keyword phrase "fish supply" yielded pages related not to over fishing, but to food and other supplies for fish in aquariums. "Drug resistant infections", the term from the investment survey summarized in Table 1, didn't have measurable searches, but the shorter phrase "drug resistant" turns up 95 searches per week.

Table 5: Keyword Searches Related to Global Environmental Management Issues from Table 1

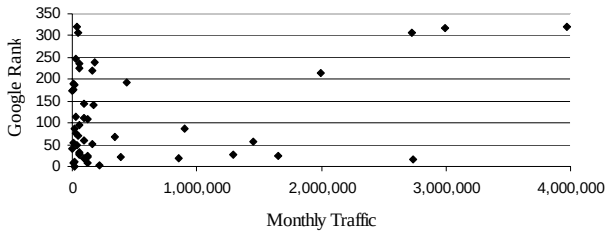
Survey Rank	Issue	Estimated Weekly Searches
1	Global Warming	18,386
2	Solar Power	5,801
	Wind Power	3,107
	Ethanol Fuel	745
	Total Searches Related to Oil Dependency	9,653
3	Hunger	2,768
	Malnutrition	1,497
	Total Searches Related to Hunger and Malnutrition	4,265
4	Air Pollution	4,372
5	Water Pollution	3,565
6	Over Fishing	65
7	Epidemic	940
8	Drug resistant	95
9	Waste Disposal	2,135

There were many searches on "waste disposal", but one suspects that many of these were not related to environmental management. The more specific and next most used phrase related to "hazardous waste disposal" generated an estimate of 1,333 searches. The search statistics for "solar power", "wind power" and "ethanol fuel" are combined in Table 5 to represent the number two investment survey category of "oil dependency". The individual search statistics for "hunger" and "malnutrition" are also combined to represent the third investment survey category of "hunger and malnutrition". The result is that the rank order of the first five categories from the investment survey in Table 1 matches exactly with the rank order of the key word search as reported in Table 5.

Statistical Analysis of Web Page Traffic and Google Search Order

A regression analysis relating the Google search rank order to the independent variable of monthly traffic statistics revealed an intuitively contradictory result that traffic rank is strongly correlated with the numerical value of the Google search rank, indicating that sites with more monthly traffic appear lower on the list. The regression results are reported in Table 6, but the results are evident by looking at the underlying data reported previously in Table 2 where many of the pages near the bottom 350 of the Google search had the highest monthly traffic. This is also illustrated by Figure 1.

Figure 1: Google Rank Against Monthly Traffic



One potential statistical problem with this analysis is that the dependent variable is a ranking, representing ordinal data For example, the R Square from the regression is a relatively low 0.097 while the independent variable "monthly traffic" is significant with a p-value well below the standard 0.05. The rank order character of the dependent variable (ordinal) may somewhat distort the statistical results of the regression equation.

Table 6: Regression Analysis of Google Search Order Explained by Monthly Traffic

Dependent Variable: Google Search Rank *Independent Variable: Monthly Traffic As Estimated by TrafficEstimate.com* R Square = .097		
Variable	*Coefficient*	*p-Value*
Intercept	97.065	0.38
Monthly Traffic	0.0000341	0.0291*
N = 49	* Statistically significant	

A statistical tool developed to deal with this problem is an ordinal regression, as explained by software provider SPSS: "Ordinal regression allows you to model the dependence of a polytomous ordinal

response on a set of predictors, which can be factors or covariates. The design of ordinal regression is based on the methodology of McCullagh [11] . . . "

In this case, the numeric data of monthly traffic are treated as covariates, a modified logit regression model is then estimated, translating each numeric covariate explanatory variable into an estimate of the ordinal rank of the dependent ordinal variable.

As the results in Table 7 indicate, the correlation between the monthly traffic and the Google search rank is similar to the regression results, but the significance level is only about 0.066, or slightly less than the 95 percent confidence level. The results of the ordinal regression also include a table (too lengthy to reproduce) converting the Google search rank to traffic estimates and providing a significance estimate for each rank.

Table 7: Ordinal Regression of Google Search Order Explained by Monthly Traffic

	Chi-Square	Significance
Final Model	3.374	0.066
Link function: Logit		

Statistical Analysis of Keyword Search Statistics and Investment Banking Survey Rank Order

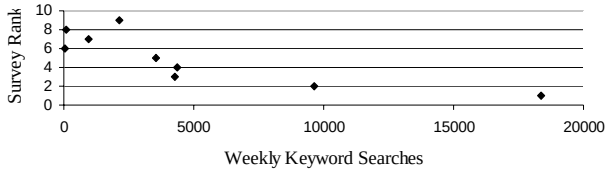
A second regression equation, summarized in Table 8, was used to estimate the how the rank order of the investment survey could be estimated using the keyword search statistics as the independent variable.

Table 8: Regression Analysis of Investment Survey Rank as Explained by Weekly Keyword Search

Dependent Variable: Investment Survey Rank as Reported in Table 1 *Independent Variable: Weekly Keyword Search Statistics as Estimated by Webfooted.net and Reported in Table 5* R Square = .66		
Variable	*Coefficient*	*p-Value*
Intercept	6.827	0.00004**
Weeklyly Searches	-0.00038	0.00784**
N = 9 ** Statistically significant at 0.01 level		

The regression analysis can be interpreted as showing that the rank order works well from about rank 7 and higher. The value of the intercept, about 7, is very significant as is the coefficient on the weekly searches variable, indicating that as searches go up the rank number goes down. Figure 2 illustrates the data used in the regression analysis.

Figure 2: Innvestment Survey Rank Against Weekly Keyword Search Statistics (Data from Table 5)



One approach suggested by a reviewer of this paper that could be used to correct for the ordinal data problem associated with the ordinary least squares regression was to convert the data from the keyword search statistics into to a rank order. A Kendall's Tau B statistic was then calculated with the results, summarized in the Table 9 below, indicating that the rank order from the investment survey and the keyword search statistics rank order are significant at about the 0.01 level.

Table 9: Kendall's Tau B of Investment Survey Rank with Keyword Search Statistics Rank

		Survey Rank
Rank of Keyword Search Statistics	Correlation Coefficient	0.704 **
	Significance (two-tailed)	0.009
** Correlation is significant at the 0.01 level (two-tailed)		

An ordinal regression was also applied to this data with results summarized in Table 10 confirming the highly significant relationship between the rank from the investment survey and the keyword search statistics related to each survey topic, the data summarized in Table 5. The number of weekly keyword searches is the covariate in this model. The significance for the final model, including the constant term was 0.000, as reported in Table 10 near the top. As indicated by the regression results, the keyword search covariate was significant at the 0.017 level, slightly less than that for the ordinary least squares regression and with a similar coefficient.

The negative value of the estimated coefficient for the keyword search covariate (-0.001 as reported in Table 10) is roughly comparable to the result from the ordinary least squares regression and indicates that topics with a higher weekly keyword search rate have a smaller number in the ranking scale. In other

words, the number one ranked site would have the most searches; the number two site would have the next highest, and so on.

Table 10: Ordinal Regression of Investment Survey Rank Explained by Keyword Search Statistics (Data from Table 5)

Model	Chi-Square	df	Sig.
Final	16.902	1	.000
Parameter Estimates			
		Estimate	Sig.
Threshold	SurveyRank = 1	-17.653	.183
	SurveyRank = 2	-9.333	.049
	SurveyRank = 3	-6.052	.011
	SurveyRank = 4	-4.682	.025
	SurveyRank = 5	-2.879	.074
	SurveyRank = 6	-1.433	.256
	SurveyRank = 7	-0.514	.662
	SurveyRank = 8	0.465	.716
Location	KeywordSearch	-0.001	.017

CONCLUSIONS

Although the monthly traffic tended to rise for web pages farther down in the Google search result, it is very unlikely that this pattern would continue for the 2 million pages resulting from the search on "global environmental management." Nevertheless, search engines such as Google that give a higher weight to hyperlink structure than to traffic may not be properly prioritizing web pages for some kinds of searches. For example, a researcher looking for freely available data on the global environment, would have to patiently sift through at least 350 of the top pages from the Google site to find one of the most visited, and arguably one of the most useful web pages. A serious researcher would have to skip through many pages such as introductory text from Wikipedia to find the very useful data sets available from NOAA that are obviously in high use given the traffic estimates.

The hyperlink structure of the Google search, with its focus on how a "random surfer" would find their way through the internet using hyperlinks, hearkens back to the days when the article describing this technology was written [1], back in the 1990s when many of us were using random hyperlinks because good search tools were just becoming available.

Keyword search statistics are generally available on the internet to provide information that can help to design web pages, but this data may also be useful in more general ways, beyond the choice of catch phrases used to attract traffic. The keyword search data provides a measure of the content that internet users are looking for. As such, the data can guide the decisions related to web page content or even for more general issues such as business investment or market research.

The research presented in this paper suggests that the keyword search data may be usefully substituted for more expensive survey research, opening up a variety of research opportunities that might have been prohibitively expensive otherwise. Google has a number of initiatives underway to provide keyword search data, not available during the research phase of this paper but currently available as "Google Adwords", easily found by doing a search on Google Adwords [6]. The Google website provides a breakdown of keyword search data by month. It is provided for Google clients or prospective clients who may be interested in buying keywords to improve their standing in search results, another example of Google monetizing some aspect of internet search.

Future research in this area may focus on applying the keyword search data to other topics or on a more thorough examination of issues related to global environmental management. In any event, it is encouraging to see all the efforts that are underway to deal with the problems of the global environment. Some of the tools that have been developed to manage the network environment can also be used to help manage the physical environment.

REFERENCES

1. Brin, S. and L. Page (1998). The anatomy of a large scale hypertextual web search engine. *Proceedings of the 7th International Conference on the World Wide Web*, 107-117
2. Dahan, Ely and John R. Hauser (2002). The virtual customer. *Journal of Product Innovation Management*. 19(5), 332-334.
3. Dhyani, Devanshu; Wee Keong Ng, and Sourav S. Bhowmick (2002). A survey of web metrics, *ACM Computing Surveys*. 34(4), 469-503.
4. Egghe, L., & R. Rousseau (1990). *Introduction to Informetrics*, Elsevier Science Publishers, Amsterdam.
5. Eyob, Ephrem (2006). E-Commerce transactions: an empirical analysis & understanding of web-based applications. I*ssues in Information Systems*, 7(2), 192-196.

6. Google (2007). Available at: adwords.google.com
7. Keen, Cherie; Wetzels, Martin; de Ruyter, Ko; and Feinbery, Richard (2004). E-tailers versus retailers: Which factors determine consumer preferences? *Journal of Business Research*, 57, 685-695.
8. Korgaonkar, Pradeep K; & Wolin, Lori D., (1999). A multivariate analysis of web usage. *Journal of Advertising Research*, 39(2), 53-68
9. Lo, Bruce W. N. & Sedhain, Rosy Sharma (2006). How reliable are website rankings? Implications for e-business advertising and internet search. *Issues in Information Systems*, 7(2), 233- 238
10. Liechty, John; Ramaswamy, Venkatram; & Cohen, Steven Ho (2001). Choice menus for mass customization: An experimental approach for analyzing customer demand with an application to a web-based information service. *Journal of Marketing Research*, 38(2), 183-196.
11. McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.
12. NOAA (National Oceanic and Atmospheric Administration) Satellite and Information Service, Available at: www.nesdis.noaa.gov.
13. Norusis, Marija (2007). Bivariate correlation, *SPSS Guide to Data Analysis*. SPSS.
14. Shi, Yuquan (2006). The search engine visibility of Queensland visitor information centres' websites. *Issues in Information Systems*, 7(2), 228-232.
15. Taylor, Chris (2007). "Be Rich. Go Green", *Business 2.0 Magazine*, January 27.
16. Wolfers, Justin & Zitzewitz, Eric (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107-126.
17. Yuwono, B. & Lee, D. (1996). Search and ranking algorithms for location resources on the World Wide Web. *Proceedings of the 12th International Conference on Data Engineering*, March, 164-171.