https://doi.org/10.48009/2_iis_2008_111-116

USING GOOGLE KEYWORD STATISTICS TO EXPLAIN CHANGES IN TRAFFIC TO INTERNET SITES RELATED TO GLOBAL ENVIRONMENTAL MANAGEMENT

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

ABSTRACT

Google recently provided a keyword statistics tool to support its "AdWords" program that conducts a deep evaluation of individual web sites. It provides detailed information on the keywords in the site and an index of monthly and average statistics for these keywords over the past year for users of the Google search engine. The data is intended to help with the design of internet sites and to support bidding for keywords that when purchased become AdWords that draw traffic through the sponsored links of Google search. This paper describes an analysis of the traffic, keywords, and search rank results for a sample of internet sites related to global environmental management for March 2007 and March 2008. Surprisingly, the analysis shows absolutely no correlation between changes in traffic and changes in keyword statistics. There is only weak evidence that the Google search rank affects traffic. One interesting result was that highly trafficked sites received less traffic while less trafficked sites received more traffic, a regression to mean traffic.

Keywords: Traffic analysis, Google, Search engine visibility, Keyword statistics, Global environment

INTRODUCTION

A relatively new source of revenue for Google is the "AdWords" program, which allows advertisers to bid on keywords that will drive traffic to their site from the sponsored links that appear at the top and to the right of a keyword search using Google. To help with the bidding process and to provide ideas for keywords, Google provides indexes of the keyword searches from internet users relying on their search engine. These statistics can also be used by anyone to help understand what internet users are searching for and what keywords may draw traffic to a site. Over the past few years, as the program was developed, the data has been alternatively freely available or available only to clients. As of March 2008, indexes summarizing the data are freely available at: https://adwords.google.com/select/KeywordToolExternal

The option for this tool that allows evaluation of individual web sites provides an impressive display of the actual Google search algorithm. The tool finds all of the significant keywords within the entire site and provides a monthly and average index of the number of searches for these keywords by users of the Google search engine. As the Google site explains, the keyword tool is intended to provide keyword ideas. At the top of the page, Google suggests, "It is All About Results" (trademarked phrase). Near the top of the page Google also provides the disclaimer that: "We cannot guarantee that these keywords will improve your campaign performance." This phrase might seem like a standard but unnecessary legal disclaimer given the widely held perception of the importance of Google keyword search as a method for finding information on the internet.

The goal of this research was to verify and quantify the relationship between search statistics and internet traffic. In addition, since keyword statistics ultimately affect page ranking from Google search, the relationship between traffic and Google search rank is investigated.

LITERATURE REVIEW

No research could be uncovered relating keyword statistics to aggregate traffic for a large number of sites not engaged in Google advertising programs. The literature draws heavily on surveys of users or log files from one or a few individual servers [6, 7, 8]. Also, there are studies quantifying how many users rely on search to find web sites, about 85 percent [2, 13].

The hyperlink relationship of keywords purchased through the AdWords program to site traffic is clear. When an internet user types a word or words into the Google search engine, the sponsored links that appear at the top or to the right of the search results are prioritized with the AdWords program. A significant amount of literature is devoted to optimizing the bidding process and selection of keywords, AdWords when you pay for them [3, 11, 12].

Since many internet users relay on search results to navigate the internet, managers of sites that do not

rely on advertising are often interested in increasing their traffic by adjusting content. Improving their sites ranking in the search engines is likely to bring more visibility.

The Google search algorithm relies on a weighted average of hyperlink structure and site keywords to prioritize pages [1, 17]. One effort to game this algorithm has been to set up multiple other pages that link to a target page [16], but Google has responded by creating a ranking of hyperlinks. For example, a page with a link from a page that itself has many links will get a higher priority than a page with a link from a page with a few links. Still, as Lo and Sedhain demonstrate, web page rankings are very sensitive to the method of ranking [9]. Interestingly, one significant source of data, network traffic, is not used to prioritize page rankings. Some researchers cite this as a shortcoming of the Google algorithm [16]. One study demonstrates that traffic actually seems to be negatively correlated with page rank so that more trafficked pages appear lower in the search [15]. The actual Google algorithm uses over 200 different factors which are "hard to infer" [4], but Evans shows that PageRank as determined by keywords and hypertext plays a statistically significant role, though not an overriding role, in determining search order. Also, domain age seems to play a role with the idea that older sites could be more trusted.

The Google keyword tool provides an impressive amount of data. This source of data might meet some of the data requirements discussed by Eyob [5] who suggests that future research on customer satisfaction could incorporate more explanatory variables. It might also contribute to the often laborious task of website content analysis [6, 10, 18]. Shi [14] also suggests that organizations should consider what data can be collected to "optimize their search engine marketing strategies.

DATA AND RESEARCH METHODOLOGY

Data for traffic, keyword frequencies, and Google page rank were collected in March in 2007 and 2008 to provide the basis for the changes over a one-year period. The phrase "global environmental management" was entered into the Google search engine at the beginning of the month for each year. The top 350 results from March 2007 were saved for analysis. Traffic for the preceding 30 days was estimated using the tool at trafficestimation.com.

One limitation of the traffic estimation tool is that traffic can only be estimated for the home page of the site. For example, the Wikipedia page related to the environment ranks relatively high on the page ranking but only traffic through the home page could be estimated. Since this would seriously distort the traffic important to this analysis, only search results that linked to home pages were used for the analysis. Another restriction was that a few of the search results were related to other topics such as the global environment for management of information technology. These pages were also eliminated. Also, for unknown reasons the traffic estimator could not generate a number for a few of the pages in 2007 or 2008 and these pages were also not included. As a result, a sample of 41 sites qualified for the analysis.

One limitation to the Google keyword statistics is that they are reported as an index rather than as raw keyword data. Although the indexes should be expected to accurately represent relative proportions of keyword searches, the actual data for the average index for the year seems often to be the same number as the index for the most recent month. This may be the result of a flaw in the data reporting.

Table 1 reports representative data for the top ten trafficked sites in 2007 illustrate the data that was collected. Both traffic and keyword searches went down over the one year period. Of the sites ranked by Google in 2007, 41.5 percent feel out of the Google search in 2008. Only four of the sites increased their ranking.

Average traffic to sites fell by 5021 visits from 2007 to 2008. The average percentage change of visits per site actually increased by 24.3 percent because many of the less trafficked sites had large increases in traffic. Heavily trafficked sites declined in traffic, but the percentage losses were not so large. The index for total keyword searches fell by an average 19.9 percent.

One likely explanation for the decline in keyword searches is that near the end of January 2007 the Intergovernmental Panel on Climate Change released a widely publicized report concluding that global warming would likely be far more destructive than previously thought. The topics were covered widely by the news media and probably encouraged many internet users to search on related topics.

Regression analysis of the data using the software SPSS was the primary research tool used to tests the following two research hypothesis:

Table 1: Data for Top Ten Trafficked Sites from 2007

2007 Traffic Rank	URL	2007 Monthly Traffic	Traffic Change 2007 to 2008	Keyword Statistics Change Feb. 2007 to 2008	2007 Google Rank	2008 Google Rank
1	www.nesdis.noaa.gov	3,968,000	-1,802,500	-20.87	319	325
2	teachearth.com	2,990,000	N.A.	-29.16	317	Below 850.
3	chge.med.harvard.edu	2,740,500	-878,200	-50.3100	15	839
4	sustainabilityscience.org	2,724,400	-862,100	-32.1400	306	188
5	www.erb.umich.edu	1,994,100	-269,000	-42.2300	214	56
6	www.sage.wisc.edu	1,657,700	-270,200	-27.1100	24	223
7	na.unep.net/	1,449,200	-457,900	-45.9500	58	Below 850
8	environment.newscientist .com	1,291,600	-581,200	-48.8600	28	Below 850
9	www.ceage.vt.edu/	898,400	100,700	-1.0200	87	Below 850.
10	nigec.ucdavis.edu/	851,100	N.A.	N.A.	20	N.A.
Based on a sample of 350 pages from Google in 2007. N.A. indicates that data was not available for 2008, the 10th ranked site was down.						

HA1: Changes in the number of overall searches for keywords at an internet site will be correlated with the traffic at the site.

HA2: Changes in the Goggle search rank for an internet site will be correlated with the traffic at the site.

RESUTLS

Table 2 reports the regression results obtained using the change in traffic at each of the 41 internet sites from March 2007 to March 2008 as the dependent variable. Independent variables are the change in keyword statistics and whether or not the site was able to stay within the Google page rankings in 2008, coded as 1 that they were still found by the search and 0 if not. The results are surprising.

The high p-Value for the relationship between traffic change and keyword change suggests about a 92 percent chance that there is no correlation, not even a hint of a relationship. For companies that stayed on the Google search results in both 2007 and 2008, the p-Value for the variable is significant at just about the 0.05 level, but the sign is negative. This means that sites who were included in both the 2007 and 2008 Google search experienced a decline in traffic as opposed to sites that fell off the search in 2008.

Table 2: Regression Analysis of Monthly Traffic Explained by Change in Keyword Search Statistics and Staying on Google Rankings

Dependent Variable: Change in Traffic 2007 to 2008 for Each Internet Site R Square = .084		
Variable	*Coefficient*	*p-Value*
Constant	23572.46	0.879
Keyword Change	390.84	0.915
On Google Results	-204667.16	0.070*
N = 41 * Statistically significant at .1 level		

The following tables present the results of some alternative model specifications in a data exploration effort to see how well these results hold up. In table 3 the variable "On Google Results" coded 0 and 1 in table 2 is revised to change in Google Rank. Since Google only reports the first 850 searches in a routine search, sites that fell out of the rankings are coded as arbitrarily as a rank of 1000 in 2008. The results are similar to table 3 and the same general conclusions can be drawn. However, coding the information this

manner marginally reduces the explanatory power of the Google search rank variable.

Table 3: Regression Analysis of Monthly Traffic Explained by Change in Keyword Search Statistics and Change in Google Rankings

Dependent Variable: Change in Traffic 2007 to 2008 for Each Internet Site R Square = .074		
Variable	*Coefficient*	*p-Value*
Constant	254975.68	0.110
Keyword Change	565.91	0.879
Google Rank Change	-253.95	0.090*
N = 41 * Statistically significant at 0.1 level		

With the idea that there may be a lag between the changes in traffic and the change in keyword statistics, table 4 examines a one month lag between keyword search and traffic. Perhaps internet users find a new site and then start to visit it more often over the next 30 days. As the results indicate, in the table 4 specification the keyword change still appears to be unrelated to traffic with a p-Value slightly above 0.8. The variable for staying on the Google search results for the year is slightly more significant, the sign is still negative, indicating a benefit from being dropped by the Google search.

Table 4: Regression Analysis of Monthly Traffic Explained by Change in Keyword Search Statistics from Previous Month (Jan.) and Staying on Google Rankings

Dependent Variable: Change in Traffic 2007 to 2008 for Each Internet Site R Square = .085		
Variable	*Coefficient*	*p-Value*
Constant	67743.94	0.632
Keyword Change Lagged One Month	-855.74	0.809
On Google Results	-204216.87	.068
N = 41 * Statistically significant at 0.1 level		

In table 5 the variables for change in traffic and keyword statistics are expressed as percentages, resulting in a relationship where although the coefficients are still statistically insignificant, at least they have the expected signs: both positive.

Table 5: Regression Analysis of Percentage Change in Monthly Traffic Explained by Percentage Change in Keyword Search Statistics from Previous Month (Jan.) and Staying on Google Rankings

Dependent Variable: Percentage Change in Traffic 2007 to 2008 for Each Internet Site R Square = .044		
Variable	*Coefficient*	*p-Value*
Constant	2.674	0.485
Percentage Keyword Change Lagged One Month	5.227	0.498
On Google Results	1.147	.263
N = 41 * Statistically significant at 0.1 level		

The most significant equation that could be derived from the data is reported in table 6, relating the percentage change in traffic over the year to the sites ranking in terms of traffic in 2007 and staying on the Google search results for both years. Smaller sites tended to gain traffic while larger sites tended to lose, significant with a p-Value below 0.05. Also, staying on the Google search results tended to increase traffic with a p-Value of slightly above 0.1 (about a 90 percent chance of a relationship).

Table 6: Regression Analysis of Percentage Change in Monthly Traffic Explained by the Traffic Rank from 2007 and Staying on Google Rankings

Dependent Variable: Percentage Change in Traffic 2007 to 2008 for Each Internet Site R Square = .195		
Variable	*Coefficient*	*p-Value*
Constant	-2.353	0.034*
2007 Traffic Rank	0.096	0.009*
On Google Results	1.534	0.110
N = 41 * Statistically significant at 0.05 level		

A final model is presented in table 7 that adds to the variables from table 6 the percentage change in keyword statistics lagged by one month. Although the impact of the percentage change in keyword statistics is still not statistically significant, at least the coefficient has the expected positive sign. Also, even though the correlation among the independent variables is relatively low (none higher than .15) the

model is a little unstable over the variety of specifications.

Table 7: Regression Analysis of Percentage Change in Monthly Traffic Explained by the Percentage Change in Keyword Search Statistics, the Traffic Rank from 2007, and Staying on Google Rankings

Dependent Variable: Percentage Change in Traffic 2007 to 2008 for Each Internet Site R Square = .196		
Variable	*Coefficient*	*p-Value*
Constant	-2.124	0.166
Percentage Keyword Change	0.007	0.827
2007 Traffic Rank	0.096	0.009*
On Google Results	1.544	0.113
N = 41 * Statistically significant at 0.05 level		

CONCLUSIONS

The first research hypothesis that changes in keyword search statistics will be correlated with changes in traffic is clearly not supported. This is a surprising result. One original goal of this research was to calibrate to relationship between searches and traffic, but the no effect null hypothesis must be accepted in this case. Further research on different types of searches might help to clarify this relationship.

The second research hypothesis that changes in Google search rank will be correlated with changes in traffic is weakly supported (at about the 89 percent confidence level) by models examining percentage changes in traffic and keywords. It is also surprising that a stronger result was not identified. As with the first hypothesis, further research on a broader sample might be warranted.

Since the results of this research were somewhat surprising, a number of model specifications were reported in the data analysis section as evidence that the weak or negative results did not seem to be a result of model misspecification. None of the model specifications came close to supporting the idea that changes in Google keyword statistics were related to changes in traffic. Google has provided a very sophisticated design tool with its keyword statistics analysis that is intended to help fine tune web content. At least for non-sponsored search results, use of the keywords does not seem to increase traffic. The Google disclaimer related to the keyword statistics tool "*We cannot guarantee that these keywords will improve your campaign performance,"* turns out to be a surprisingly good piece of advice.

For the data collected in this study, the variable that proved to have the strongest role in predicting changes in traffic over the one-year period was the rank of the page in terms of traffic in 2007. More trafficked sites tended to lose traffic while less trafficked sites tended to gain. This may be a result of a commonly observed statistical tendency of many types of data to regress to the mean value of the population. The least square tool used in this analysis was tagged as "regression" in part because it was used in early studies to uncover this tendency.

REFERENCES

1.Brin, S. and L. Page (1998). "The anatomy of a large scale hypertextual web search engine," *Proceedings of the 7th World Wide Web Conference.*

2. Dreze, Xavier and Fred Zufryden (2004). "Measurement of online visibility and its impact on Internet Traffic," *Journal of Interactive Marketing*" Winter 18(1), pp. 20 – 37.

3. Dwihananto, Dimas; Moh, Teng-Sheng (2007). "Effectively finding the right keywords for the target audience," *Signal Processing and Information Technology, 2007 IEEE International Symposium*, pp. 766-771.

4. Evans, Michael P. (2007). "Analyzing Google rankings through search engine optimization data," *Internet Research*, 17(1), pp. 21 – 37.

5. Eyob, Ephrem (2006). "E-Commerce transactions: an empirical analysis & understanding of web-based applications," I*ssues in Information Systems*, Volume VII, No. 2, pp. 192-196.

6. Huang, Eugenia Y. (2006) "Is revamping you web site worthwhile," *Industrial Management and Data Systems*, 105(6), pp. 737 – 751.

7. Korgaonkar, Pradeep K; & Wolin, Lori D., (1999). A multivariate analysis of web usage. *Journal of Advertising Research*, Vol. 39. No. 2, pp. 53-68

8. 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*, Vol. 38, No. 2, pp. 183-196.

9. Lo, Bruce W. N. and Sedhain, Rosy Sharma (2006). "How reliable are website rankings? Implications for e-business advertising and internet search.," *Issues in Information Systems*, 7(2), pp. 233- 238

10. Lo, BWN, E. Claire, P. Gong (2005). "Cultural impact on the design of e-commerce websites: part 1 – site formation and layout," *Issues in Information Systems* 6(2), pp. 182 – 188.

11. Mehta, Aranyak; Saberi, Amin; Vazirani, Umesh; Vazirani, Vijay (2007). "AdWords and generalized online matching," *Journal of the ACM*. October 54(5)

12. Rusmevichientong, Paat andDavid P. Williamson (2006). "An adaptive algorithm for selecting profitable keywords for search-based advertising services," *Proceedings of the 7th ACM conference on Electronic Commerce.* Pp 260 – 269.

13. Schmidt-Mnaz, Nadine and Gaul, Wolfgang (2004) "Measurement of online visibility," *Operations Research Proceedings 2003*. Springer, pp. 205 - 212

14. Shi, Yuquan (2006). The search engine visibility of Queensland visitor information centres' websites. *Issues in Information Systems*, Vol. VII, No. 2, 228-232.

15. Webb, G. Kent (2007). "Analysis of web pages and metrics related to global environmental management", *Issues in Information Systems*, 7(2). pp. 7 – 13.

16. Weinman, Joe (2007). "A New Approach to Search", *Business Communications Review*. October, pp. 19 – 29.

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.

18. Zhao, JJ. & Zhao, S.Y. (2004). "Internet technologies used by IC. 500 corporate web sites," *Issues in Information Systems*. (4), pp. 366 – 372.