ICIC Express Letters
Part B: Applications
Volume **7**, Number **9**, September **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **2029–2034**

A STUDY ON SEARCH ENGINE RANKING FACTORS – AN EXAMPLE OF GOOGLE SEARCH ENGINE

Lily Lin[1], Shuang-Kai Tsao[2] and Huey-Ming Lee[2]

[1]Department of International Business
China University of Technology
No. 56, Sec. 3, Hsing-Lung Road, Taipei 116, Taiwan
lily@cute.edu.tw

[2]Department of Information Management
Chinese Culture University
No. 55, Hwa-Kung Road, Yang-Ming-Shan, Taipei 11114, Taiwan
mikekai.tw@gmail.com; hmlee@faculty.pccu.edu.tw

Received February 2016; accepted May 2016

Abstract. *In the study we apply multiple regression analysis to finding out the Google search engine ranking factors. We describe the essential criteria of search engine ranking. The result will be helpful for users to let their websites be shown on Google search top. We choose one set of 20 specific keywords related to one kind of product to get the websites ranking of each keyword on the first three pages of Google as our samples. The result shows URL length, backlink anchor text, keyword appearing in h1 tag, low related keywords and external links which show significant difference, as the major factors to affect the website ranking. To prove the model is applicable on other products, we choose another product to do multiple regression analysis again, and the result is the same.*
Keywords: Search engine optimization (SEO), Multiple regression analysis, Search engine ranking

1. **Introduction.** As we know that the Internet marketing has matured, the role of search engines has become more important. Thus, SEO (Search Engine Optimization) [1] can help the website to feature prominently on search engines like Google. It is important for Internet marketing to get the website or page to rank on the first page of search engine that can help improve website traffic and sales.

When the user needs some information, search engines are the most commonly used tools. Evans [3] selected seven of the most influential factors from Google's search engine including: number of pages indexed, page rank of a website, number of in-links, domain age, DMOZ directory submissions, yahoo directory submissions and Del.icio.us bookmarks. SEO is a kind of search rule of the search engine, and it is optimized by the structure, the key word, the title, links of the website, etc. [5]. Google [4] issued SEO Starter Guide, which refers Title, Meta description, Anchor text, Image Alt structure, etc., are very important factors. Though many references as above discussions are about search engine ranking factors, we would like to use a model to decide the most important factors in this study. Thus, even the factors will change in the future, we can use the proposed model to do analysis again to find out the factors.

2. **Preliminaries.** Google considers over 200 factors that have an influence upon the website ranking. Based on Dean [2], we ignore the factors only can be retrieved from website owners, such as average sessions duration, bounce rate, pageviews/sessions. In this study, we only choose the external factors as shown in Figure 1, from which we can get the related data sources. The total numbers of external factors which we can get are

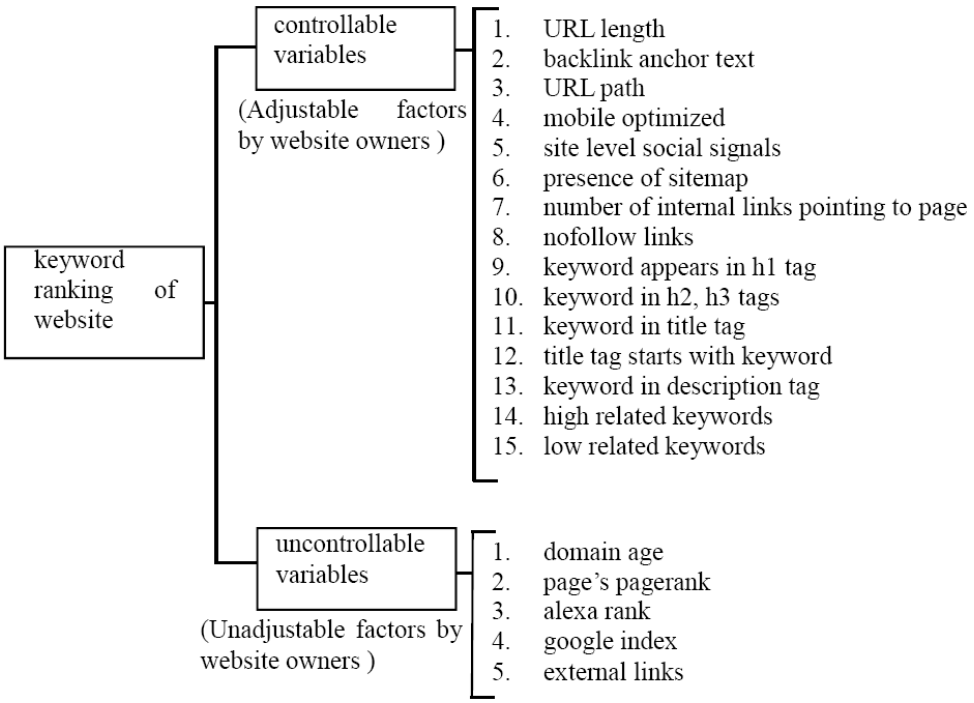

FIGURE 1. External factors

20 factors divided into controllable variables (adjustable factors by website owners) and uncontrollable variables (unadjustable factors by website owners) as Figure 1.

Two hypotheses proposed show as follows:

H_0^c: controllable variables without influence upon keyword ranking of website;

H_1^c: controllable variables with an influence upon keyword ranking of website;

and

H_0^{uc}: uncontrollable variables without influence upon keyword ranking of website;

H_1^{uc}: uncontrollable variables with an influence upon keyword ranking of website;

H_0: null hypothesis;

H_1: alternative hypothesis;

c: controllable variables.

3. **The Proposed Model.** We use the multiple regression analysis method and the regression model [6] shown as follows:

$$Y_{1*1} = \beta_{n*1} X_{1*n} + \varepsilon_{1*1} \tag{1}$$

Y: ranking of website

X: independent variables, $[X = X_0, X_1, \ldots, X_{20}]$

β: vector of unknown parameters $\beta_i = [\beta_0, \beta_1, \ldots, \beta_{20}]$

ε: error term, $\varepsilon \sim \mathrm{N}(0, \sigma^2)$

$$\begin{aligned} H_0 &: \ \beta_i = 0 \\ H_1 &: \ \beta_i \neq 0 \quad (i = 1, 2, \ldots, 20) \end{aligned} \tag{2}$$

Regression analysis generates Equation (1) to describe the statistical relationship between predictor variables (X) and the response variable (Y). The data source of dependent and independent variables, methods of measurement and data attributes are shown in Table 1.

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that we can reject the null hypothesis (H_0). In other words, an independent variable that has a low p-value is likely to be a meaningful

TABLE 1. Methods of measurement of variables

Variable symbol	**Variable name**	**Data source**	**Methods of measurement**	**Data attribute**
Y	website ranking	Google search engine	website ranking	continuous variable
Independent variable (X) controllable variables				
X_1	URL length	website	Microsoft Office Excel (SUM)	continuous variable
X_2	backlink anchor text	website	extract data from website	dummy variable
X_3	URL path	website	extract data from website	continuous variable
X_4	mobile optimized	website	extract data from website	dummy variable
X_5	site level social signals	website	extract data from website	dummy variable
X_6	presence of sitemap	website	extract data from website	dummy variable
X_7	number of internal links pointing to page	website	extract data from https://goo.gl/fST6P8	continuous variable
X_8	nofollow links	website	extract data from https://goo.gl/fST6P8	continuous variable
X_9	keyword appearing in h1 tag	website	extract data from website source code	dummy variable
X_{10}	keyword in h2 tags	website	extract data from website source code	dummy variable
X_{11}	keyword in title tag	website	extract data from website source code	dummy variable
X_{12}	title tag starts with keyword	website	extract data from website source code	dummy variable
X_{13}	keyword in description tag	website	extract data from website source code	dummy variable
X_{14}	high related keywords	website	extract data from website source code	continuous variable
X_{15}	low related keywords	website	extract data from website source code	continuous variable
Independent variable (X) uncontrollable variables				
X_{16}	domain age	website	Microsoft Office Excel (SUM)	continuous variable
X_{17}	page's pagerank	website	extract data from https://goo.gl/jrT57X	continuous variable
X_{18}	alexa rank	website	extract data from http://goo.gl/Zo1Ao9	continuous variable
X_{19}	google index	website	Google Search Engine	continuous variable
X_{20}	external links	website	extract data from https://goo.gl/fST6P8	continuous variable

addition to the model because changes in the independent variables are related to changes in the response variable.

To choose the keywords for model, we set the monthly average searches (average the number of searches for the term over 12-month period in 2015) range to be $8,000 \sim 13,000$. In the "Competition" column, we choose the competition for a keyword to be medium or high. Based on above two rules, through Google AdWords we pick up 20 keywords of "苦茶油" related, and the websites on the first 3 pages for those 20 keywords, 600 websites are our samples to do analysis by performing a multiple regression analysis in SPSS Statistics. The output shows in Table 2. Moreover, to check whether the result of other product keywords is consistent, we choose 20 keywords of "發熱衣" related, another 600 websites to do analysis. According to the result in Equations (3) and (4), all of coefficient signs of independent variables are the same in these two equations.

TABLE 2. Expected correlation of independent variables

Variable symbol	Variable name	Expected correlation	Beta distribution (Coefficient)	p-value	Symbol correlation coreect/wrong
X_1	URL length	Positive	0.236	0.022	correct
X_2	backlink anchor text	Negative	−0.247	0.004	correct
X_3	URL path	Negative	−0.076	0.437	correct
X_4	mobile optimized	Negative	0.102	0.223	wrong
X_5	site level social signals	Negative	−0.50	0.637	correct
X_6	presence of sitemap	Negative	0.207	0.014	wrong
X_7	number of internal links pointing to page	Negative	0.006	0.934	wrong
X_8	nofollow links	Positive	0.301	0.019	correct
X_9	keyword appearing in h1 tag	Negative	−0.185	0.054	correct
X_{10}	keyword in h2 tags	Negative	−0.116	0.257	correct
X_{11}	keyword in title tag	Negative	−0.028	0.921	correct
X_{12}	title tag starts with keyword	Negative	0.074	0.780	wrong
X_{13}	keyword in description tag	Negative	0.158	0.147	wrong
X_{14}	high related keywords	Negative	0.139	0.100	wrong
X_{15}	low related keywords	Positive	0.099	0.209	correct
X_{16}	domain age	Negative	0.009	0.929	wrong
X_{17}	page's page rank	Negative	0.096	0.417	wrong
X_{18}	alexa rank	Positive	0.007	0.933	correct
X_{19}	google index	Negative	0.241	0.449	wrong
X_{20}	external links	Negative	−0.226	0.110	correct

4. **Implementing Results.** We implement the keywords of "苦茶油" and "發熱衣" in our proposed model, and we have the following results.

(1) keywords of "苦茶油".

F is 4.53 and a low R-Squared (0.351) does not affect the interpretation of significant variables. The result equation shows as follows:

$$\begin{aligned} Y = {} & 0.236X_1 - 0.247X_2 - 0.076X_3 + 0.102X_4 - 0.50X_5 + 0.207X_6 + 0.006X_7 \\ & + 0.301X_8 - 0.185X_9 - 0.116X_{10} - 0.028X_{11} + 0.074X_{12} + 0.158X_{13} \\ & + 0.139X_{14} + 0.099X_{15} + 0.009X_{16} + 0.096X_{17} + 0.007X_{18} + 0.241X_{19} \\ & - 0.226X_{20} + 1.886 \end{aligned} \tag{3}$$

(2) keywords of "發熱衣".

F is 4.91 and a low R-Squared (0.366) does not affect the interpretation of significant variables. The result equation shows as follows:

$$\begin{aligned} Y = {} & 0.225X_1 - 0.213X_2 - 0.082X_3 + 0.119X_4 - 0.35X_5 + 0.219X_6 + 0.004X_7 \\ & + 0.321X_8 - 0.175X_9 - 0.136X_{10} - 0.034X_{11} + 0.061X_{12} + 0.125X_{13} \\ & + 0.210X_{14} + 0.131X_{15} + 0.011X_{16} + 0.088X_{17} + 0.007X_{18} + 0.290X_{19} \\ & - 0.284X_{20} + 1.814 \end{aligned} \tag{4}$$

Some coefficient signs of independent variables do not match prior expectation, and we only choose those factors matching prior expectation to analyze. There are total 10 variables matching prior expectation (URL length, backlink anchor text, URL path, site level social signals, nofollow links, keyword appearing in h1 tag, keyword in h2 tags, low related keywords, alexa rank and external links).

To find the best model, we use the method of forward selection stepwise regression to choose the most appropriate variables from 10 variables to do analysis by performing a multiple regression analysis in SPSS Statistics. The result of model can build a regression equation as Equations (5) and (6). The output results are shown as Table 3 and Table 4, we can see that all of the predictor variables of X_1, X_2, X_9, X_{15}, and X_{20} are significant because all of their p-values are less than the common alpha level of 0.05. However, the p-value for X_{10} (0.091) is greater than the common alpha level of 0.05, which indicates that it is not statistically significant.

(1) keyword of "苦荼油":

$$Y = 0.336X_1 - 0.179X_2 - 0.091X_9 - 0.102X_{10} + 0.188X_{15} - 0.214X_{20} + 1.012 \quad (5)$$

(2) keyword of "發熱衣":

$$Y = 0.312X_1 - 0.210X_2 - 0.212X_9 - 0.113X_{10} + 0.294X_{15} - 0.185X_{20} + 1.001 \quad (6)$$

Table 3. The results of stepwise regression analysis for keywords of "苦荼油"

Variable name	URL length	backlink anchor text	keyword appearing in h1 tag	keyword in h2 tags	low related keywords	external links
Beta Distribution (Coefficient)	0.336	−0.179	−0.091	−0.102	0.188	−0.214
(t)	2.528	−1.861	−1.416	−1.008	1.859	−1.483
p-value	0.007	0.001	0.005	0.091	0.006	0.031
F Value = 6.49			Adj R-Sq = 0.688			

Table 4. The result of stepwise regression analysis for keywords of "發熱衣"

Variable name	URL length	backlink anchor text	keyword appearing in h1 tag	keyword in h2 tags	low related keywords	external links
Beta Distribution (Coefficient)	0.312	−0.210	−0.212	−0.113	0.294	−0.185
(t)	2.419	−2.128	−0.194	−0.216	1.678	−1.337
p-value	0.004	0.001	0.009	0.100	0.012	0.019
F Value = 6.21			Adj R-Sq = 0.611			

5. **Conclusions.** In Table 3 and Table 4, the final results show URL length, backlink anchor text, keyword appearing in h1 tag, low related keywords and external links which show significant difference, as the major factors to affect the website ranking. Some factors such as average sessions duration, bounce rate, and pageviews/sessions, are also very important factors for website ranking. However, in this study we ignore those factors data since they can be retrieved only from website owners.

Acknowledgment. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Search engine optimization, *Wikipedia, the Free Encyclopedia*, https://en.wikipedia.org/wiki/Search_engine_optimization.
[2] B. Dean, *Google's 200 Ranking Factors: The Complete List*, http://www.backlinko.com/google-ranking-factors, 2016.
[3] M. P. Evans, Analysing Google rankings through search engine optimization data, *Internet Research*, vol.17, no.1, pp.21-37, 2007.
[4] Google, *Search Engine Optimization Starter Guide*, https://static.googleusercontent.com/media/www.google.com/zh-TW//webmasters/docs/search-engine-optimization-starter-guide.pdf, 2011.
[5] C. J. Luh and L. H. Liao, *Google Search Engine Ranking and the Weight of Factors*, http://oplab.im.ntu.edu.tw/csimweb/system/application/views/files/ICIM/20110094, 2011 (in Chinese).
[6] P. S. Mann, *Statistics for Business and Economics*, John Wiley & Sons, Inc., New York, 1995.