# A Linear Regression Model for Assessing the Ranking of Web Sites Based on Number of Visits

Dowming Yeh, Pei-Chen Sun, and Jia-Wen Lee

National Kaoshiung Normal University
Kaoshiung, Taiwan 802, Republic of China
{dmyeh, sun}@nknucc.nknu.edu.tw

**Abstract.** Many web sites are designed and established without sufficient professional skills and resources. The quality of these websites is often dubious. Therefore, how to evaluate the quality of a web site has become an important issue. In this work, the stepwise regression method is applied to assess the ranking of web sites for two different categories. The ranking is based on the average number of visits per day. A total of fourteen factors frequently found in the literature are considered as independent variables in developing the model. The regression analysis result shows that the regression models differ for two different categories of web sites, but there are three variables common to the two resulting models, Latest update, Broken links, and Help features. The average prediction accuracies of both models exceed 75%.

## 1 Introduction

Constructing a web site are nowadays straightforward, demanding few technological efforts. Besides professional developers, there are more and more amateurs engaged in the production of many web sites [1]. Although the number of sites flourishes extensively, the content and the quality of some sites do not improve simultaneously. Therefore, how to evaluate the quality of a web site has become an important issue [4].

A web site should be evaluated from both the content and the design of the site [5]. The content covers the content of an entire site as well as an individual web page. The characteristics of the content for the entire site include not only the information content, but also support for transaction and elements of entertainment [1]. The content of a web page comprises of elements of text, color, graphics, mages, audio, animation, and so on. The criteria for assessing the content include correctness, periodical update, completeness, organization and clearness, attractiveness, value [3].

The design of a web site consists of the interface or the layout design of web pages and the navigational design of web sites. It could be evaluated in four aspects [6]: usability, functionality, reliability, and efficiency. Usability concerns how to assist users in using the site effectively. Functionality includes search and navigational mechanisms. Reliability addresses the correctness of the link and the errors incurred

N. Koch, P. Fraternali, and M. Wirsing (Eds.): ICWE 2004, LNCS 3140, pp. 45–49, 2004.

by different configurations. Efficiency tackles factors that might affect the download speed of web pages and the accessibility of the information on a page. Ivory and Hearst evaluate web pages with 11 measures covering both the content and the design aspects [2]. In this work, the stepwise regression method is applied to assess the ranking of web sites for two different categories in Taiwan. The ranking is based on the average number of visits per day. Such ranking represents somehow the ranking of user satisfaction, and therefore the quality of a web site.

## 2 Research Variables

An indicator of the ranking of web site may be based on how often the site is utilized, which reflects true opinions of real users somewhat. Therefore, we base the dependent variable on the average number of visits per day to a web site during a span of three months. Because the number of visits varies enormously for top sites and poor sites, we map these numbers into their ranking positions instead of the using the actual numbers. The definition of the variable follows:

$$Y_i = 100 - \frac{100}{(n-1)} \times (i-1)$$

$Y_i$ is the score of the $i$-th ranked in the set of sites, and $n$ is the number of web sites in the set under study.

The possible values for independent values are all scaled to the range from 0 to 100 for easier manipulation and exploration of the model. There are a total of fourteen independent variables under consideration as follows:

1. Site map: A numerical function $X_1$ is defined to be 100, if the site provides a site map or a TOC; 50, if the site provides a navigational menu; 0, if none is provided.
2. Help feature: The corresponding function $X_2$ is given as 100, if providing organized help feature such as FAQ; 80, providing interactive help feature such as bulletin board; 60, providing email and responding within a week; 40, providing email and responding more than a week; 0, providing none.
3. Latest update: We define its function $X_3$ to be 100, updated within 3 days; 75, updated within a month; 50, updated within 3 months; 25, updated within a year; 0, otherwise (including no update indication).
4. Font count: Function $X_4$ is defined as $100 - 20 \times |n - 3|$, where n is the number of fonts and $1 \leqq n \leqq 7$; 0, otherwise.
5. Color count: Function $X_5$ is defined to be $100 - 20 \times |n - 7|$, where n is the number of colors and $3 \leqq n \leqq 11$; 0, otherwise.
6. Foreign Language Support: The corresponding function $X_6$ is defined to be 100, if providing both English and Simplified Chinese versions; 80, if providing English or Simplified Chinese version, plus another version for other language; 60, if providing English or Simplified Chinese version; 0, otherwise.
7. Search mechanisms: Function $X_7$ is given as 100, if providing search mechanism covering the entire site; 50, if providing search mechanism covering part of the site (bulletin board, for example); 0, if providing none.

8. Link count: The corresponding function $X_8$ is defined as in [6].
9. Scrolling: Let $n$ be the ratio of the length of the page divided by the length of the screen. The function $X_9$ is 100, if $n \leqq 3$; $100 - 15\times (n - 3)$, if $4 \leqq n \leqq 6$; $55 - 10 \times (n - 6)$, if $7 \leqq n \leqq 8$; $35 - 5 \times (n - 8)$, if $9 \leqq n \leqq 10$; 0, otherwise.
10. Word count: We define the variable $X_{10}$ as ($n$ is the total word count) 100, if $651 \leqq n \leqq 1300$; $100 - 20 \times \text{ceiling}((n - 1300) / 650)$, if $1301 \leqq n \leqq 3900$; 80, if $326 \leqq n \leqq 650$; 60, if $160 \leqq n \leqq 325$; 40, if $n < 160$; 0, otherwise.
11. Broken link: We adopt the function proposed by Olsina et al. in [6] as $X_{11}$.
12. Static page size: Function $X_{12}$ is defined as (let the size be $n$) 100, if $n \leqq 80$ KB; 90, if 80 KB $< n \leqq 100$ KB; $100 - 10 \times \text{truncate}(n / 50 \text{ KB})$, if 100 KB $< n \leqq$ 500 KB; 0, otherwise.
13. Image label: The function proposed by [6] is used as $X_{13}$.
14. Number of panes regarding frames: Function $X_{14}$ is the same as that Olsina et al. propose [6].

Our initial regression model therefore takes the following from:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + b_6 X_6 + b_7 X_7 + b_8 X_8 + b_9 X_9 + b_{10} X_{10} + b_{11} X_{11} + b_{12} X_{12} + b_{13} X_{13} + b_{14} X_{14}$$

## 3 Data Analysis

The source data of this research comes from a web site called HotRank (http://www.hotrank.com.tw). HotRank maintains visit information for several thousand web sites. Theses web sites are further classified into different categories from which two categories, *academy and literature* and *shopping* are chosen. The sites that appear in the seasonal ranking list from January to March 2003 are chosen as sample data.

To ensure the uniformity of the visit data, we further requires that the chosen sites are operational since January 1st 2003. The number of sites in the academy and literature category conforming to the condition stated above is 305. However, all sites ranked after the 90th position are eliminated because their average number of visits per day falls to zero and other sites failing to provide the average number of visits per day are also taken out. After reduction, there are a total of 60 sites from the academy and literature category. Similarly for the shopping category, the number of sites reduces to 198. We randomly select 30 sites respectively from two categories to conduct regression analysis. Some of the data for the independent variables result from average values of the homepage and other nine pages randomly selected from the site.

Before the stepwise regression analysis, the Pearson analysis is performed and the results indicate that all the interdependency values between any two independent variables are less than 0.55, so there is no collinear relationship between any two of the fourteen independent variables. Therefore, it is appropriate to consider all these variables in the stepwise regression analysis. The Durbin-Watson (DW) test is employed to examine that the error $e_i$ of independent variables should not be self-related. From our analysis, the DW values of the models for the academy and literature category and shopping category are 1.476 and 1.589, respectively. This reveals that there

is no significant self-relation in the two models and the prediction of the models can be trustworthy. Results of the stepwise regression analysis are described in the following paragraphs.

F test examines the overall regression model, also called as Analysis of Variance (ANOVA). The F values of the analysis result are 20.324 and 20.637 respectively for the academy and the shopping categories. Both of the P-Values are less than 0.005. Therefore, the linear relationships of our models are well established.

The coefficient of multiple determinations measures the proportion that independent variables are able to explicate the dependent variable. Another related measure is the adjusted coefficient of determination, which is considered more representative than the coefficient of determination. Applying these measures, the coefficient of multiple determination and adjusted coefficient of determination for the model of the academy and literature category are 0.765 and 0.727, respectively. As for the shopping category, the coefficient of multiple determinations and adjusted coefficient of determination are 0.768 and 0.730, respectively. This indicates that the independent variables in the models can account for around 73 % of the dependent variable.

Finally, the t test is to examine whether there is a significant linear relationship between every independent variable and the dependent variable. If there is no significant linear relationship between an independent variable *X*i and the dependent variable *Y*, the coefficient of the variable $b_i$ should be set to zero. Applying the t test to all the fourteen independent variables, the result shows that the final regression model of the academy and literature category is

$$Y = -58.308 + 0.493 \times \text{Last\_update} + 0.472 \times \text{Broken\_links} + 0.219 \times \text{Site\_map} + 0.475 \times \text{Help\_feature}$$

And the model for the shooping category is

$$Y = -46.345 + 0.494 \times \text{Help\_feature} + 0.371 \times \text{Static\_page\_Size} + 0.378 \times \text{Last\_update} + 0.259 \times \text{Broken\_links}$$

The adjusted coefficients of determination of the two models are 0.727 and 0.730, respectively. Considering the diversities of web sites, such degree of accounting precision is well acceptable.

In order to check the applicability of our model, another 15 sample sites are randomly selected for each category from the set of sites that is not selected previously in establishing the models. We apply the model to predict the rankings of these sites and compare them with the actual rankings. The result shows that the average prediction accuracy of the model for the academy and literature category is 76.0 %, and that of the model for the shopping category is 79.2 %.

## 4 Discussions and Conclusions

The difference of the independent variables in the two models suggests that there are indeed different indicators for different categories of web sites and the effects of these indicators are also different. There are three variables common to the two resulting models, Latest update, Broken links, and Help features. These three variables address different aspects of a site, information, reliability and usability, respectively. This

implies a good web site must excel in various aspects. Latest update involves the content of the web site and more specifically how often the content of the site is updated. In a fast changing world, this is certainly a great concern to the site users. Broken links address the reliability issue as well as the correctness of the content in a web site. It is always frustrating for users to chase after a link only to find it leads to nowhere. Help features, on the other hand, tackle the usability of a web site. As the a web site evolve in functionality and complexity, usability issues commonly found in software applications surface inevitably and online help plays an important part in usability of a software system.

The Site map variable address yet another important aspect of a web site, i.e., navigation issues. The reason that it is not present in the model of the shopping category is that the content of sites in this category map directly to a hierarchical structure reflecting the structure of product catalogue. However, the static page size is a significant factor for these online shopping sites. A page with an immense number of bytes would take time to download, which is not contradictory to the efficiency that most online shopper expects. While we may imagine users in the sites of the academy and literature category to be more leisurely when they surf these sites, thus efficiency is not a great concern.

We apply linear regression method for assessing the ranking of a web site based on the number of visits per day. A total of fourteen factors frequently found in the literature are considered as independent variables in developing the model. The regression analysis result shows that the regression models differ for two different categories of web sites. The average prediction accuracy of the model for the academy and literature category is 76.0%, and that of the model for the shopping category is 79.2%.

## References

1. Huizingh, E. K. R. E., "The content and design of web sitesan empirical study," *Information & Management* 37, 2000, pp.123-134.
2. M.Y. Ivory, R.R. Sinha, and M.A. Hearst, "Empirically Validated Web Page Design Metrics," *Proc. Conf. Human Factors in Computing Systems*, vol. 1, ACM Press, New York, Mar. 2001, pp. 53-60.
3. Katerattanakul, P. and Siau, S., "Measuring information quality of web sites: development of an instrument, " Proceeding of the 20th international conference on Information Systems, January 1999, pp.279-285.
4. Lin, J. C., and Lu, H., "Towards an understanding of the behavioral intention to use a web site," *International Journal of Information and management* 20, 2000, pp.197-208.
5. McMurdo, G., "Evaluation web information and design," *Journal of Information Science*, 24(3), 1998, pp.192-204.
6. Olsina, L., Godoy, D., Lafuente, G. J., and Rossi, G., "Quality Characteristics and Attributes for Academic Web Sites," *Proc. Of Web Engineering Workshop at WWW8*, Toronto, Canada, 1999.