Study of patterns in the hyperlink structure of large sites

Ivan Blekanov
Saint-Petersburg State University
7-9 Universitetskaya Naberezhnaya,
St. Petersburg, 199034, Russian
+7 921 339 53 43
i.blekanov@gmail.com

Sergei Sergeev
Saint-Petersburg State University
7-9 Universitetskaya Naberezhnaya,
St. Petersburg, 199034, Russian
+7 921 381 23 62
slsergeev@yandex.ru

Evgenii Klemeshov
Saint-Petersburg State University
7-9 Universitetskaya Naberezhnaya,
St. Petersburg, 199034, Russian
+7 961 808 63 52
zheklem14@gmail.com

ABSTRACT

This paper presents experimental results on websites of four universities investigated with the aim of drawing up lists of pages and links. Variants of functions approximating the experimental data are considered. The resulting approximation estimates indicate that the approximating functions proposed are of high precision.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory – g*raph algorithms.*

General Terms

Measurement, Experimentation.

Keywords

Web-site, approximation, hyperlinked structure of the site, webometrics, web-graph.

1. INTRODUCTION

Today, you can hardly come across an organization that does not have its own website. In a sense, a website is the face of an organization and the quality is the site is of great importance. The quality of a website consists of many components – page count, page design, page content, and site structure.

A website structure is usually presented in the form of a directed graph whose nodes are documents and whose edges are links connecting these documents [1]. Many researchers today study the global structural characteristics of information networks on the Web. For example, Broder A. and Kumar F. [1] represented large website communities in the form of a strongly connected graph component, components In, Out and Tubes; [2]-[5] deal with distribution of external links and citation index of various university sites; authors in [6] introduce site connectivity characteristics; [7] is dedicated to the study of the method of automatic classification of links and pages by their characteristics; and others.

This paper aims at studying a website structure. More precisely, it tries to identify the kind of functional dependence that exists between the page count of a site and the internal link count of that site.

2. HYPOTHESIS VERIFICATION

2.1 Problem statement

Special crawler [8] was used to solve this problem. In scanning the site, the robot creates two lists: a list of found pages (web graph nodes) and a list of links connecting the found pages (web graph edges).

An act of finding *e* links will be considered as a step of the scanning algorithm. That is, $E_i = e \cdot i$ links are found after *i* steps. Let us denote with $v_i = v(E_i)$ the number of pages found after *i* steps. We denote the number of all the links and the number of all pages found in the site with e_0 and $v(e_0) = v_0$ respectively. The search robot can find not only the number of pages and links in the site, but also get a graph of the $v(e)$ function.

Obviously, $v \le e$. Here, $v_0 < e_0$ (it is very rare when $v_0 = e_0$ for websites). It is also obvious that increase in $v(e)$ is gradually slowed down. This is due to the fact that with increasing number of found pages, the likelihood that the next link will point to a page already found increases.

Let us try and choose an analytic function approximately describing experimental function $v(E)$, with the following properties:

1. Passes through the origin in the plane (e, v).
2. Increases with an increase in e_0.
3. Derivative of the function decreases.
4. The function has a simple form and is dependent on a small number of parameters.

*IWAIT'*15, Oct. 8–10, 2015, Aizu-Wakamatsu, Japan.

2.2 Experiment

Let's compare the function with experimental set $V_i, E_i\ (i = \overline{1,N})$, where $N = \left\lceil \frac{e_0}{e} \right\rceil$. We investigated the websites of four universities and obtained the following sets (Table 1).

To assess the approximation quality, we will use average relative error:

$$\Delta = \frac{1}{N}\sum_{1}^{N}\frac{|V_i - v_i|}{v_i}.$$

Table 1. Universities sites studied

University	URL	Total pages	Total links
Saint-Petersburg State University	www.spbu.ru	*41183*	*2664000*
Moscow State University	www.msu.ru	*47832*	*1891000*
The University of Aizu	www.u-aizu.ac.jp	*4161*	*49900*
The University of Tokyo	www.u-tokyo.ac.jp	*> 17000*	*240000*

Next, we consider three possible approximating functions:

$$v^{(1)} = \alpha \cdot E^{\beta},$$

$$v^{(2)} = \frac{\alpha \cdot E}{E + \beta},$$

$$v^{(3)} = \alpha \cdot (\ln(1+E))^{\beta}.$$

Let us examine them one by one:

1) We take the logarithm of the equation:

$$\ln v^{(1)} = \ln\alpha + \beta \cdot \ln E.$$

This gives a system of linear equations

$$x + A_i \cdot y = B_i,\ i = \overline{1,N},$$

where $x = \ln\alpha,\ y = \beta,\ A_i = \ln E_i,\ B_i = \ln V_i$.

Its solution with method of least squares:

$$\alpha = \exp\frac{C_1C_4 - C_2C_3}{N\cdot(C_1^2 - C_3)},\quad \beta = \frac{C_1C_2 - C_4}{C_1^2 - C_3},$$

where

$$C_1 = \sum_1^N A_i, C_2 = \sum_1^N B_i, C_3 = \sum_1^N A_i^2, C_4 = \sum_1^N A_iB_i.$$

2) System of linear equations:

$$\alpha \cdot E_i - \beta \cdot V_i = V_i \cdot E_i,\ i = \overline{1,N}.$$

Its solution with method of least squares:

$$\alpha = \frac{a_3 \cdot a_4 - a_2 \cdot a_5}{a_1 \cdot a_4 - a_2^2},\quad \beta = \frac{a_2 \cdot a_3 - a_1 \cdot a_5}{a_1 \cdot a_4 - a_2^2}$$

Where

$$a_1 = \sum_1^N E_i^2,\ a_2 = \sum_1^N E_iV_i,\ a_3 = \sum_1^N E_i^2V_i,$$

$$a_4 = \sum_1^N V_i^2,\quad a_5 = \sum_1^N V_i^2E_i.$$

3) We take the logarithm of $v^{(3)}$:

$$\ln v^{(3)} = \ln\alpha + \beta \cdot \ln\ln(1+E)$$

We get a system that almost coincides with the first case. Its solution will be the same, except that in the first case $A_i = \ln E_i$, while in this case,

$$A_i = \ln\ln(1+E_i).$$

2.3 Results of the experiment

Figures 1, 2, 3 and 4 present the graphs of functions $v^{(1)}, v^{(2)}, v^{(3)}$ and V, for each of the universities listed.

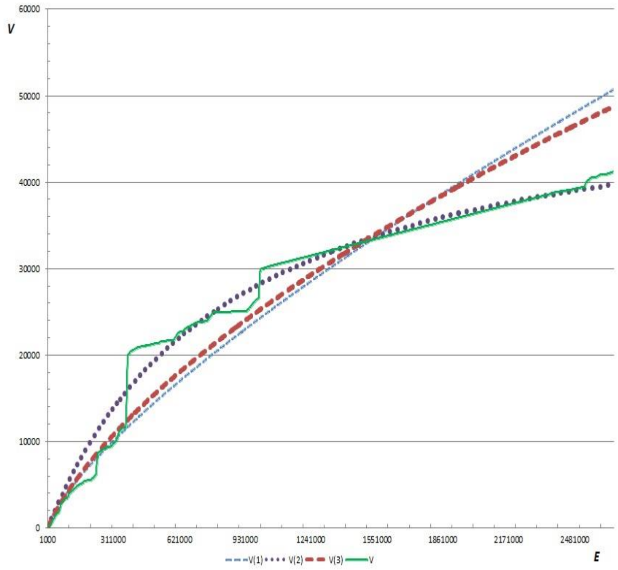

Figure 1. Functions $v^{(1)}, v^{(2)}, v^{(3)}$ and V for Saint-Petersburg State University

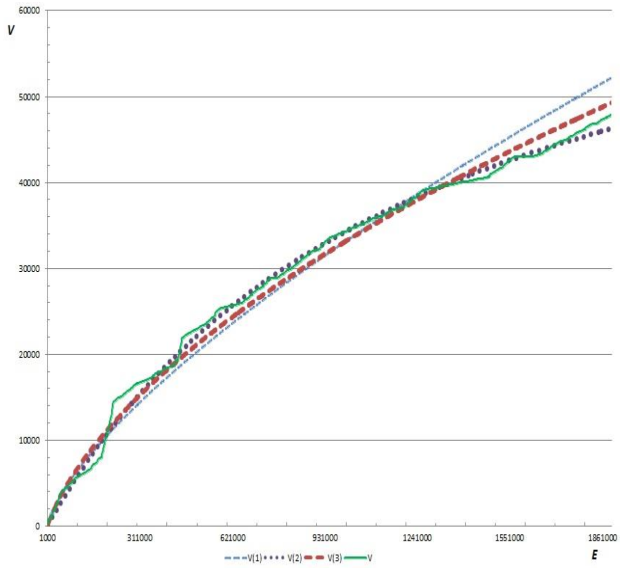

Figure 2 . Functions $v^{(1)}, v^{(2)}, v^{(3)}$ and V for Moscow State University

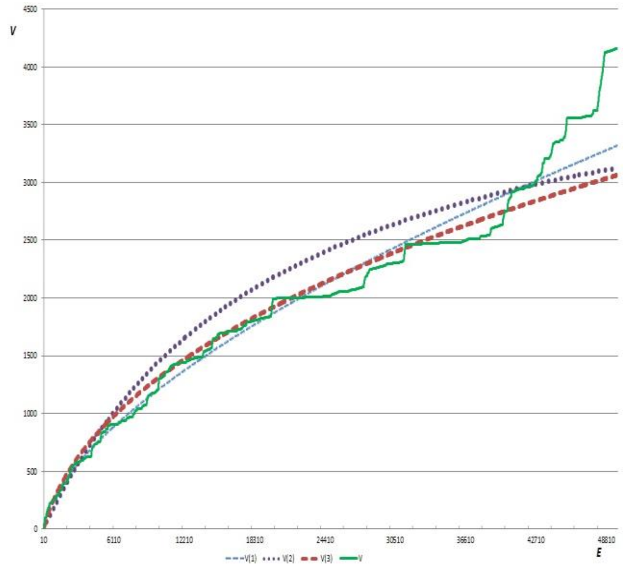

Figure 3. Functions $v^{(1)}, v^{(2)}, v^{(3)}$ and V for the University of Aizu

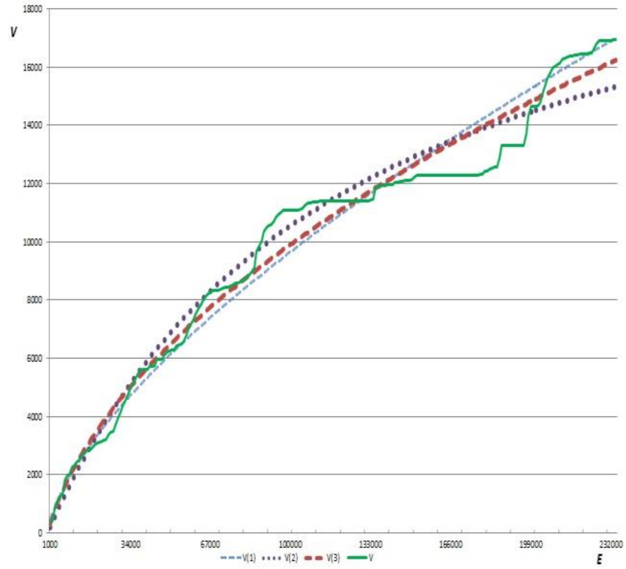

Figure 4. Functions $v^{(1)}, v^{(2)}, v^{(3)}$ and V for the University of Tokyo

Table 2 shows the relative errors of each of the formulas for each of the universities.

Table 2. Average relative error of functions $v^{(1)}, v^{(2)}, v^{(3)}$ for each university

University	$v^{(1)}$	$v^{(2)}$	$v^{(3)}$
Saint-Petersburg State University	0,166	0,141	0,099
Moscow State University	0,092	0,037	0,014
The University of Aizu	0.167	0.209	0.229
The University of Tokyo	0.068	0.076	0.062

3. CONCLUSION

The paper considered the problem of finding a function that approximates the experimental graph of dependence of web page count on link count. Three approximations were proposed. It was revealed that the best approximations are obtained for power and linear fractional function. The functions proposed can be used to study the parameters of sites and their clustering.

4. ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, grant № 15-01-06105

5. REFERENCES

[1] Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, 2000. Graph structure in the Web: Experiments and models. In WWW9 (Vol. 33, #1–6), Elsevier Science, pp: 309–320.

[2] Thelwall, M., 2013. Webometrics and Social Web Research Methods. University of Wolverhampton, pp: 8-39.

[3] Thelwall, M., Zuccala, A. 2008. A university-centred European Union link analysis. Scientometrics, 75(3), 407-420.

[4] Thelwall, M., Wilkinson, D., Musgrove, P. B., 2005. National and international university departmental web site interlinking, part 2: Link patterns. Scientometrics, 64(2): 187-208

[5] Pechnikov, A. and A. Nwohiri, 2012. Webometric analysis of Nigerian university websites. Webology. Vol. 9, Num. 1, June. (http://www.webology.org/2012/v9n1/a95.html).

[6] Blekanov, I.S., S.L. Sergeev and A.I. Maksimov, 2014. Analysis of the topology of large Web segments using Broder’s bow-tie model. Life Science Journal, Vol. 11: 258-261.

[7] Kenekayoro, P., K. Buckley, M. Thelwall, 2014. Automatic classification of academic web page types. Journal Scientometrics, Vol. 101(2): 1015-1026.

[8] Blekanov I., S. Sergeev and I. Martynenko, 2012. Construction of subject-oriented web crawlers using a generalized kernel. Scientific and technical bulletins of St. Petersburg State Polytechnic University. St. Petersburg State Polytechnic University, # 5 (157). pp: 9-15.