An Algorithm to Approximate the Total Number of Site Pages Using a Portion of Its Structure

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

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

Aleksei Maksimov
Saint-Petersburg State University
7-9 Universitetskaya Naberezhnaya,
St. Petersburg, 199034, Russia
+7 911 212 76 76
nightsnaker@gmail.com

ABSTRACT
This article considers the algorithm that allows to measure the webpages and hyperlinks number after the fractional inspection based on the set of equations that correlates the web-crawled and found webpages number. The experiment results that verify algorithm's performance capability were pointed out within.

Categories and Subject Descriptors
G.2.2 [Discrete Mathematics]: Graph Theory – graph algorithms.

General Terms
Measurement, Experimentation.

Keywords
Web-site, hyperlinked structure of the site, webometrics, web-graph, site size determination.

1. INTRODUCTION
One of the main tasks of webometrics is the massive websites inspection to be split into two – deep webpages analysis and its' interaction with web-space [1]. This article generally considers the websites' different attributes as inherent properties. The website structure is commonly accepted (i.e. [2]) to be presented as a direct graph therefore it contains the vertices as webpages, the edges as hyperlinks, connectivity, length etc. In the second place – the ergonomic factors such as website design and usability. Third is the content attributes – the main topic, tags etc. Considering these factors is the key for the preset website selection. The evaluation of these factors is possible after inspecting, for example, the one tenth of the total number of website hyperlinks (that is comparable to the proper book selection – in order to understand the book is worth reading you are to look through the part of it). But to understand what part of the website was inspected already we must possess the number of the webpages (graph vertices). Nowadays there is only one known way to determine the true number of the webpages – the whole web-graph inspection. As long as the massive organizations' websites (i.e. universities) been estimated of tens or even hundreds of webpages with the zillions of hyperlinks its measurement is widely considered to be the resource-intensive task.

The main idea of the offered approximate webpages and hyperlinks calculation method premised on its part is as follows. At first all the hyperlinks on every webpage are to be retrieved. Then every page followed by every hyperlink is to be inspected for another set of hyperlinks etc. Some links lead to new pages, some of them are cycled in order the new-leading links share to be reduced. The relevant logic is in the works [3, 4]. Here the authors put forward the algorithm that allows to measure the website size after the inspection of its part. It is based on the new leading hyperlinks appearing slowdown.

*ICAIT'*16, Oct. 6–8, 2016, Aizu-Wakamatsu, Japan.

2. THE WEBPAGE AND HYPERLINKS NUMBER CALCULATION METHOD

2.1 Mathematical model
Let us consider the approximate procedure of a web-graph random walk.

1. Making the sets:
   a) An *NLinks* $\{a_i\}$. a_i set is the number of retrieved hyperlinks in *i* steps;
   b) An *NPage* $\{v_i\}$. v_i set is the number of webpages that retrieved hyperlinks in *i* steps lead to;
   c) An *UrlsList* set is the number of webpages' domain names that retrieved hyperlinks in *i* steps lead to.
2. Let us retrieve Δa_i hyperlinks from the website and add $a_i = a_{i-1} + \Delta a_i$ to *NLinks* set. Every retrieved hyperlink is to be collated with the *UrlsList* set. If there is no webpage that hyperlink leads to then the hyperlink joins the *UrlsList* set. The added webpages number is Δv_i. Then $v_i = v_{i-1} + \Delta v_i$ been added to the *NPage* set.

Let us denote all the website hyperlinks and webpages number by a_{total} and v_{total} respectively; the number of retrieved hyperlink by a, and the unique webpages number by v. Let us also introduce:

$$a_{remain} = a_{total} - a \text{ and } v_{remain} = v_{total} - v.$$

Let us denote the unique website pages set by V_{total} and still unretrieved webpages number by V_{remain}. Let us break the webpages set into subsets by the incoming hyperlinks. Let us denote the set of the webpages with n incoming hyperlinks by S_n^0 and the subset of webpages within V_{remain}, with n incoming hyperlinks by $S_n(v_{remain})$.

Then $s_n^0 = |S_n^0|$, $s_n(v_{remain}) = |S_n(v_{remain})|$.
The probability of the next retrieved hyperlink leads to the webpage from the $S_n(v_{remain})$ set equals to:

$$p_n(v_{remain}) = \frac{nS_n(v_{remain})}{a_{remain}} = \frac{nS_n(v_{remain})}{a_{total} - a}$$

Let any step has da as the number of hyperlinks retrieved. Then the expected value of the increment $s_n(v_{remain})$:

$$ds_n(v_{remain}) = -\frac{ns_n(v_{remain})}{a_{total} - a}da$$

Let us integrate between 0 and a. As far as $v_{remain} = v_{total}$ with $a = 0$, then:

$$\ln s_n(v_{remain}) - \ln s_n^0 = n(\ln(a_{total} - a) - \ln a_{total})$$

Therefore:

$$s_n(v_{remain}) = s_n^0\left(\frac{a_{total} - a}{a_{total}}\right)^n$$

Let us use the trivial par:

$$\sum_n s_n(v_{remain}) = v_{remain} = v_{total} - v$$

Therefore:

$$v = v_{total} - \sum_n s_n^0\left(\frac{a_{total}-a}{a_{total}}\right)^n \quad (1)$$

It is known (i.e. [5, 6]) that in large networks nodes are distributed by the incoming hyperlinks number follows the law

$$s_n^0 = \frac{\gamma v_{total}}{n^q} \quad (2),$$

with $\frac{1}{\gamma} = \sum_n \frac{1}{n^q}$, and q as an invariable.

$$2<q<3 \quad (3)$$

Let us plug the formula (2) into the equation (1). Then the equation that shows the correlation of the websites pages number v_{total} and the website hyperlinks number a_{total} is as follows:

$$v = v_{total}\left(1 - \gamma\sum_n \frac{1}{n^q}\left(\frac{a_{total}-a}{a_{total}}\right)^n\right) \quad (4)$$

Let us perform the *k* of consequent steps of hyperlinks and related webpages (the pages these hyperlinks lead to) accidental samplings. Let a_i hyperlinks and v_i related webpages were retrieved after the *i*-step. Then the unique equation will correspond to every step. Therefore there is the set of equations that connects the retrieved hyperlinks with the related webpages total numbers v_{total} and a_{total} respectively:

$$v_i = v_{total}\left(1 - \gamma\sum_n \frac{1}{n^q}\left(\frac{a_{total}-a_i}{a_{total}}\right)^n\right), (i = 1,2,\dots,k) \quad (5)$$

2.2 The approximate solution

As far as we need to get the approximate measurement let us assume

$$\frac{1}{n^q} \approx \frac{1}{n(n+1)} \quad (6)$$

Then the limits of this distribution were defined by the inequation (3) (except *n=1*). Therefore the normalizing factor γ:

$$\frac{1}{\gamma} = \sum_n \frac{1}{n(n+1)} = 1 \quad (7)$$

Let us plug the formula (6) into the equation (5) considering (7). Therefore:

$$v_i = v_{total}\left(1 - \sum\frac{1}{n(n+1)}\left(1 - \frac{a_i}{a_{total}}\right)^n\right)$$

Let us analyze the sum

$$\sigma = \sum\frac{1}{n(n+1)}x^n$$

$$\sigma = \sum_1^N \frac{x^n}{n} - \frac{1}{x}\sum_1^N \frac{x^{n+1}}{n+1} = \sum_1^N \frac{x^n}{n} - \frac{1}{x}\sum_2^{N+1}\frac{x^n}{n}$$

As far as the massive websites *N* value reaches several thousand therefore the approximate value:

$$\sigma = -\ln(1-x) + \frac{1}{x}(\ln(1-x) + 1)$$

Therefore:

$$v_i = v_{total}\frac{\frac{a_i}{a_{total}}}{1 - \frac{a_i}{a_{total}}}\ln\frac{a_{total}}{a_i}$$

As far as situation $a_i \ll a_{total}$ is considered let us assume:

$$v = v_{total}\frac{a}{a_{total}}(\ln a_{total} - \ln a)$$

Let us set

$$\frac{a_{total}}{v_{total}} \equiv x, \qquad \ln a_{total} \equiv y$$

Therefore the equation:

$$v_i x + a_i y = -c_i, \qquad \text{где } c_i \equiv a_i \ln a_i$$

Let us solve the overspecified set of equations with the LS method. Therefore:

$$A_{11}x + A_{12}y = C_1$$
$$A_{21}x + A_{22}y = C_2,$$

with

$$A_{11} = \sum_1^k v_i^2\,,\ A_{12} = A_{21} = -\sum_1^k a_i v_i\,,\quad A_{22} = \sum_1^k a_i^2\,,$$
$$C_1 = -\sum_1^k v_i c_i\,,\ C_2 = \sum_1^k a_i c_i \quad (8)$$

Therefore

$$x = \frac{C_1A_{22} - C_2A_{12}}{A_{11}A_{22} - A_{12}^2}\,,\ y = \frac{C_2A_{11} - C_1A_{12}}{A_{11}A_{22} - A_{12}^2} \quad (9)$$

And finally

$$a^* = e^y\,,\ v^* = \frac{e^y}{x} \quad (10),$$

with a^* and v^* are approximate values of a_{total} and v_{total}.
The formulae (8)-(10) allow to approximately measure the website size. According to the assumptions made, the formula can be used only with low value of a_i.

2.3 The experiment

Authors were to perform the experiment in order to test the offered method capability using 11 universities' sites (table 1) and determine the share of the website hyperlinks number needed to measure it approximately (the total number of hyperlinks and the total number of webpages).

Table 1. Universities' sites inspected.

#	University	URL
1	Cambridge	www.cam.ac.uk
2	MIT	www.web.mit.edu
3	Oxford	www.ox.ac.uk
4	The University of Turin	www.unito.it
5	Cornell University	www.cornell.edu
6	Emory University	www.emory.edu
7	Berkeley	www.berkeley.edu
8	Pierre and Marie Curie	www.upmc.fr
9	The University of Aizu	www.u-aizu.ac.jp
10	Syracuse University	www.syr.edu
11	Penn State University	www.psu.edu

Using a web-crawler all webpages and internal hyperlinks have been collected for the each university (from table 1). Actual values of v_{total} and a_{total} were found. Since the offered method proposed is based on accidental samplings, a Fisher-Yates shuffle algorithm described in [7] has been used in order

to shuffle the every website hyperlinks set. Then a hundredth part of hyperlinks was to be picked from the shuffled set of hyperlinks, that made up v_1 and a_1. Then in order to get v_2 and a_2 two hundredth parts of links were to be picked, three hundredth parts for v_3 and a_3 and so on. The obtained values v_i and a_i had been substituted into (4) to calculate a^* and v^*. The proportion of the visited hyperlinks was approximately determined as the ratio of a_i to a^*.

2.4 The experiment results

The table 2 reports the main experiment results.

Table 2. Actual and calculated numbers of webpages and hyperlinks of universities' sites.

University	a_{total}	a^*	v_{total}	v^*	$\frac{a_k}{a_{total}}$
Cambridge	2258483	2993431	69661	114226	0,12
MIT	590474	1363978	70751	116108	0,2
Oxford	1197127	2649532	26289	60634	0,2
The University of Turin	1195984	2975332	22796	32168	0,03
Cornell University	874555	205028	20908	22760	0,03
Emory University	190119	124999	12497	13237	0,07
Berkeley	116240	253220	10544	13207	0,03
Pierre and Marie Curie	548844	154885	9142	6989	0,03
The University of Aizu	116098	78238	3993	4375	0,07
Syracuse University	184773	454720	4979	7576	0,04
Penn State University	32984	78880	1923	3958	0,26
a_{total}, v_{total} – website hyperlinks and webpages number respectively; a^*, v^* - approximate values of $a_{total}$and v_{total}; a_k – the number of hyperlinks used for the calculation					

In table 2 the values a^* and v^* have been obtained by ratio of a_k to a^*, which is equal to 0.1 that based on 2 factors. At first all of the formulae have been obtained using the assumption that $\frac{a_k}{a^*} \ll 1$. The second is if the a_i value is too low then the representativeness of the sample decreases.

Since actual values of the total hyperlinks number differ from the calculated ones, the table shows the actual share of visited links for each university website, for which theoretical values a^* and v^* have been obtained.

3. CONCLUSION

The main results of the work are:

1) The set of equations have been obtained that connects the number of web-crawled website hyperlinks with the webpages being found.
2) There was the experiment performed in order to check the possibility to measure the webpages number premised on its part using the total numbers of the fully inspected webpages and hyperlinks of 11 universities' websites. The experiment proves the worked-out website size measurement method capability after the fractional inspection. Both overestimated and underestimated assumptions (about website size) have been noticed. The relation between the approximately calculated website size value and actual webpages total number is 2.4, and 4.3 – for the hyperlinks if worst comes to worst. The results of the experiment show that there were needed to inspect 3-26% of the total hyperlinks number to approximately measure every university website size. According to the universities' websites information it is apparent that the visited websites' pages differ by 37 times and the visited hyperlinks by 68 times. Therefore the webpages and hyperlinks numbers measurement errors in 2.4 and 4.3 times respectively are considered to be acceptable. The (5) set of equation solution method and the measurement may be improved.
3) The algorithm that implements the offered method was specified. It allows to reduce the costs of approximate webpages and hyperlinks number as well as web-graph hyperlinks measurements significantly.

4. ACKNOWLEDGMENTS

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

5. REFERENCES

[1] Thelwall, M. 2013. *Webometrics and Social Web Research Methods.* University of Wolverhampton, 8-39.

[2] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata R., Tomkins, A. and Wiener, J. 2000. *Graph structure in the Web: Experiments and models.* In WWW9, Elsevier Science, 309–320.

[3] Blekanov, I., Sergeev, S. and Klemeshov, E. 2015. *Study of patterns in the hyperlink structure of large sites*. Proceedings of the International Workshop on Applications in Information Technology (IWAIT-2015). The University of Aizu Press, 58-60.

[4] Blekanov, I.S., Sergeev, S. L. and Maksimov, A. I. 2016. *A study of structural characteristics of large websites.* Vestnik of St. Petersburg State University. Series 10. Issue 1, 78-84.

[5] Barabasi, A. and Albert, R. 1999. *Emergence of scaling in random networks. Science*, Vol. 286: 509-512.

[6] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. 1999. *Trawling the Web for emerging cyber-communities.* WWW '99 Proceedings of the 8th international conference on World Wide Web: 1481-1493.

[7] Fisher, R., Yates, F. 1948. *Statistical tables for biological, agricultural and medical research (3rd ed.)*. London, Oliver & Boyd, 26–27.