Analysis of the topology of large Web segments using Broder's bow-tie model

Ivan Stanislavovich Blekanov[1], Sergei Lvovich Sergeev[1], Aleksei Iurevich Maksimov[2]

[1]St. Petersburg State University, Universitetskaya Naberezhnaya 7-9, St. Petersburg, 199034, Russian Federation
[2]St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Kronverksky prospect 49, St. Petersburg, 197101, Russian Federation

Abstract. This paper describes an experiment on creation of a topology of typical large humanitarian-oriented and natural science-oriented Web segments based on Broder's bow-tie model. The ultimate goal of the experiment is to calculate the structural characteristics of these sites introduced by the authors and evaluate the resource costs incurred. The experimental results will be useful in planning large researches aimed at obtaining statistical data on classes of sites.
[Blekanov I.S., Sergeev S. L., Maksimov A. I. **Analysis of the topology of large Web segments using Broder's bow-tie model.** *Life Sci J* 2014;11(6s):258-261] (ISSN:1097-8135). http://www.lifesciencesite.com. 48

Keywords: Hyperlinks, analysis of large Web segments, analysis of hyperlink structures, creation of web topology, web graph, Broder's bow-tie model, characteristics of sites, site connectivity measures.

Introduction

Presently, increasing number of organizations, seeking to reflect their activities on the Web, are creating separate web resources (sites and groups of sites) and caring about improving their rankings in information retrieval systems in order to accumulate a symbolic capital and increase direct sales. Search engines generally play an important role in the formation of the rankings of sites and localized web segments.

Since 2006, major search engines (for example, *Google* and the Russian *Yandex*) are using factors related to user behaviour on web resources in their ranking methods [1-3] in a new way alongside with the traditional factors [4, 5, 6], such as word frequency, number of citations and credibility of sites, update frequency of sites, and others. For instance, user actions, for example, frequency of user visits, that earlier led to improved site rankings may now lead to "pessimization" of indicators, as search engines are becoming more "social" and are taking into account not only user profile and frequency of visit to resources but also the motivation of visitors and their behaviour strategy – frequency of repeated visit, time spent on a page and site, the logic and routes of moving from one page to another, user interests, type of content consumed, and many other factors. Modern search engines collectively consider about 300-500 of such factors when ranking.

On the other hand, there is a small amount of structural characteristics of web segments [7], which significantly influence nearly all these factors. There are a considerable number of researches [4, 6-13], in which a web fragment model is constructed and is described by a relatively small set of parameters.

The authors' general area of research is the study of links between the ranking of a site and its global characteristics. Apparently, these links are different for different categories of sites, for example, humanitarian-oriented and natural science-oriented sites. To identify significant differences between categories of sites, it is necessary to investigate a large number of sites with subsequent statistical processing of the material. The authors have developed their own analytical system for webometric investigation [14, 15], which can perform this task. In view of the extremely high cost of machine resources for such investigation, machine experiment aimed at determining the cost of a complete investigation of typical sites, whose description is the subject of this paper, is of interest.

Main part

The experiment considered the task of creating a topology based on the Broder's bow-tie model [8, 11] for typical large humanitarian-oriented and natural science-oriented Web segments with subsequent calculation of their structural characteristics and assessment of resource costs incurred.

The website of the Faculty of Journalism (JF) [16] of St. Petersburg State University was chosen as the humanitarian-oriented Web segment, while the website of the Faculty of Applied Mathematics and Control Processes (APMATH) [17] of the same university was chosen as the natural science-oriented Web segment.

With the help of a web crawler kernel-based specialized assembly of an analytical system – successfully tested in studies [15] – that was developed for webometric investigations, the hyperlink structures of both sites were identified.

These structures are described as two oriented web graphs G_{apmath} and G_{jf}, in which pages are the nodes and hyperlinks are the edges connecting these pages. Broder's bow-tie model [8] was used to obtain global characteristics. Among the set of nodes of the entire graph being investigated, this model identifies four major subsets – Central kernel (strongly connected component SCC); Origination set ("In" group); Termination set ("Out" group); "tendrils" and "tubes". Kosaraju's algorithm [18] was used to find the strongly connected component. For sparse graphs (such are the web graphs being investigated), the time taken to find the strongly connected component is proportional to the sum of all their nodes and directed edges.

The experiment calculated the following characteristics of the websites:

- size of the central kernel S_{scc};
- size of "Out" group S_{out} and "In" group S_{in};
- size of "tendrils" and "tubes" S_{tubes};
- ratio of $|S_{scc}|+|S_{in}|$ to $|S_{scc}|+|S_{out}|$.

Besides, an indicator of connectivity with other pages from the kernel was introduced for each page of the central core:

$$VC(u,v)=\begin{cases}1, \ if \ \ u\to v\\ 0, \ else\end{cases}, \ \forall u,v\in S_{scc}.$$

This indicator was the basis of averaged measure of the connectivity of each page of the kernel $MAVC(u)$ and averaged measure of the connectivity of the central kernel $MAC(S_{scc})$:

$$MAVC(u)=\frac{1}{|S_{scc}|-1}\sum_{\forall v\in S_{scc},v\neq u}VC(u,v),$$

$$MAC(S_{scc})=\frac{1}{|S_{scc}|}\sum_{\forall u\in S_{scc}}MAVC(u).$$

These measures were introduced to assess the connectivity quality of each topology, obtained during the experiment.

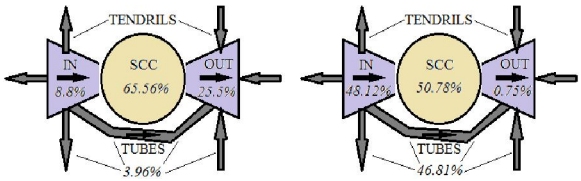
The following characteristics were used to assess the labour input of the experiment for each site investigated:

- time (T_{total}^{js} and T_{total}^{apmath}) required to create the topology and calculate the generalized characteristics specified;
- memory size ($VMem_{js}$ and $VMem_{apmath}$) required to store processed data;
- volume of Web traffic $Traffic_{js}$ and $Traffic_{apmath}$.

Outcome of experiment

The experiment was conducted on a test personal computer (PC) with processor *IntelCore i7 CPU 950 @ 3.07GHz x 8*, RAM *12 GB*, hard disk drive *250 GB*, and operating system *Ubuntu 13.10*. Data were downloaded from the Internet via a channel with a bandwidth of up to *50 Mbit/s*.

During the experiment, it was established that Web graph G_{apmath} contains *26,148* nodes and *2,025,909* directed edges, while Web graph G_{jf} has *27,361* nodes and *3,923,934* directed edges. Figures 1 and 2 show the topologies constructed for each of the graph obtained using the Broder's bow-tie model.



Figures 1 and 2: Topologies of Web graphs G_{apmath} and G_{jf} respectively

The *SCC* takes a considerable share in the component of the topology of Web graph G_{apmath} – 65.56% of the total number of pages. The "In" group takes 8.8%, the "Out" group – 25.5%, while the "tendrils" and "tubes" have 3.96% (see Figure 1). Meanwhile, these figures are approximately the same for Web graph G_{jf} (see Figure 2) (more than 46% of the total number of pages) except the "Out" components (0.75%).

Table 1 shows the numerical values of structural characteristics obtained, which were

introduced in the experiment for the analysis of the sites investigated.

Table 1. Structural characteristics of Web graphs G_{apmath} and G_{jf}.

Characteristics	Topology of APMATH site	Topology of JS site
Cardinality of the central kernel set $\lvert S_{scc}\rvert$	17,142	13,893
Cardinality of the "In" set S_{in}	2,302	13,165
Cardinality of the "Out" set S_{out}	6,669	206
Cardinality of the "tendrils" and "tubes" set S_{tubes}	1,035	12,809
Ratio $\frac{\lvert S_{scc}\rvert+\lvert S_{in}\rvert}{\lvert S_{scc}\rvert+\lvert S_{out}\rvert}$	0.82	1.92
Averaged measure of connectivity of the central kernel $MAC(S_{scc})$	0.00244	0.00151
Averaged measure of connectivity of the entire Web graph $MAC(G)$	0.00106	0.00094

The resource costs required for complete analysis of the Web segments investigated in the experiment are shown in Table 2.

Table 2. Characteristics evaluating the labour input of the experiment

Characteristics	APMATH site	JS site
Time required to build a topology and calculate structural characteristics, *min.*	$T^{apmath}_{total}=672$	$T^{js}_{total}=271$
Average time required to download one web page, *min*	0.026	0.001
Memory size on the hard disk drive for data storage, *MB*	$VMem_{apmath}=154.743$	$VMem_{js}=318.463$
Minimum RAM for data processing, *GB*	~1.5	~1.5
Average memory size required to fully process one web page, *kB*	6.06	11.92
Web traffic volume, *GB*	$Traffic_{apmath}=9.66$	$Traffic_{js}=3.1$
Average Web traffic volume for one web page, *kB*	387.29	118.13

Conclusion

The results of the experiment showed significant differences between the hyperlink structure of one of science-oriented sites and one of humanitarian-oriented sites. For example, looking at the topology of the first site (Figure 1), we can see that largest part of all the web pages are situated in the Central kernel, which is more interconnected compared to the kernel of the second site (Figure 2). Moreover, significant differences in OUT component of the first and second topologies are explained by the presence of a big number of full-text documents (PDF, DOC, DOCX, etc.). However, we can talk about great connectivity between Web graph G_{jf} components, taking the number of hyperlinks in G_{jf} and size of SCC, IN, TUBES/TENDRILS (Table 1) into account.

In order to get statistically reliable results it is necessary to research the topology of a great number of sites in each category highlighted above. Looking at the characteristic values (Table 2), which evaluate the labour of experiment held, we can estimate the cost of machine resources needed for the full investigation of any number of sites.

Corresponding Author:
Dr. Blekanov Ivan Stanislavovich
St. Petersburg State University
Universitetskaya Naberezhnaya 7-9, St. Petersburg, 199034, Russian Federation

References

1. Guha, S.K., A. Kundu and S. Bhadra, 2013. Analytical Design of Feature based Ranking. First International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013 (Vol. 10), Procedia Technology, pp: 773–780.
2. Antoniou, D., Y. Plegas, A. Tsakalidis, G. Tzimas and E. Viennas, 2012. Dynamic refinement of search engines results utilizing the user intervention. Journal of Systems and Software, Volume 85(7): 1577–1587.
3. Feuer, A., S. Savev and J.A. Aslam, 2009. Implementing and evaluating phrasal query suggestions for proximity search. Sixteenth ACM Conference on Information Knowledge and Management (CIKM 2007) (issue Vol. 34(8)), Information Systems, pp: 711–723.
4. Kleinberg, J., 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), vol. 46(5): 604-632.
5. Brin, S. and L. Page, 1998. The anatomy of a large scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7): 107-117.
6. Chakrabarti, S., 2003. Mining the Web: Analysis of hypertext and semi structured data. New York: Morgan Kaufmann, pp: 364.
7. Cho, J., and S. Roy, 2004. Impact of Web search engines on page popularity. Proceedings of the World-Wide Web Conference. Retrieved March 3, 2014 from: http://oak.cs.ucla.edu/~cho/papers/cho-bias.pdf.
8. 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.
9. Aguillo, I. F., B. Granadino, J. L. Ortega, and J. A. Prieto, 2006. Scientific research activity and communication measured with cybermetrics indicators. Journal of the American Society for Information Science and Technology, 57(10): 1296-1302.
10. Stuart, D., M. Thelwall, and G. Harries, 2007. UK academic web links and collaboration - an

exploratory study. Journal of Information Science, 33(2): 231-246.
11. Thelwall, M., 2013. Webometrics and Social Web Research Methods. University of Wolverhampton. (http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf).
12. Pechnikov, A. Webometric investigation of the websites of universities in Russia. Information Technologies, 2008, #11, pp:74-78.
13. 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).
14. 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.
15. Maksimov A.and I. Blekanov, 2013. The webometric research of the university Web segment using the Web-crawler. Control Processes and Stability: Proceedings of the 44th International Scientific Conference for Post-graduates and undergraduate students. St. Petersburg State University. pp: 403-408.
16. School of Journalism and Mass Communications of Saint Petersburg State University (Main page). Date Views March 3, 2014 www.jf.spbu.ru.
17. Faculty of Applied Mathematics and Control Processes of Saint-Petersburg State University (Main page). Date Views March 3, 2014 www.apmath.spbu.ru.
18. Sedgewick, R., 2003. Algorithms in Java, Part 5: Graph Algorithms. Addison-Wesley Professional, 3rd Edition, pp: 528.

4/16/2014