Can Link Analysis Tell Us about Web Traffic?

Marcin Sydow,
Polish-Japanese Institute of Information Technology
Koszykowa 86, Warsaw, Poland
msyd@pjwstk.edu.pl

ABSTRACT

In this paper we measure correlation between link analysis characteristics for Web pages such as in- and out-degree, PageRank and RBS with those obtained from real Web traffic analysis. Measurements made on real data from the Polish Web show that PageRank is observably but not strongly correlated with actual visits made by Web users to Web pages and that our RBS algorithm[2] is more correlated with traffic data than PageRank in some cases.

Categories and Subject Descriptors: H.3.m [Information Storage and Retrieval]: Miscellaneous

General Terms: Measurement, Algorithms, Experimentation

Keywords: Link Analysis, PageRank, RBS, Web Traffic Analysis

1. INTRODUCTION

Link analysis of the Web graph [1] proved to be a powerful tool for supporting Web searching. On the other side, Web traffic data may be treated as a measure of merit of Web documents and is of great interest to advertising companies. We measure correlation between link analysis characteristics and actual number of visits and unique users who visited Web documents. Such measurements may be regarded as a method of evaluation or comparison of ranking algorithms, or tuning their parameters. On the other hand, discovering link analysis characteristic which is highly correlated with actual Web traffic would be of great interest for researchers and commercial subjects.

2. DATA AND PREPROCESSING

We have collected 7 disjoint Polish Web subgraphs from January 2005, about 3 millions nodes each, constituting the majority of the current Polish Web. We also collected the estimated number of users and visits made to about 3 millions of unique Polish urls, from November 2004 (no fresher data was available). All this data was kindly permitted to the author by 3 Polish Internet companies [2]. All multiple, intra-document and intra-domain links were originally removed, which made the graphs extremely sparse (step 0). Further preprocessing and data unification was done by the author and concerned: removal of isolated nodes (step 1), url aggregation over the query part (since the queries were originally removed on the traffic data side)(step 2), removal of non-http protocol (since all graph side data was http-type) aggregating urls over the "fragment" part (starting with the # symbol) (step 3), graph nodes renumbering, etc. Finally, the intersection of traffic and graph urls made the experimental data sets even smaller (step 4) (see table 1). The resulting 7 data sets, $g1, ..., g7$, were much smaller Web subgraphs but were fully equipped with the traffic data. Next, for each of them we computed 4 vectors representing in-degree, out-degree, visits and users values for each page, respectively. We computed, visualized and inspected value distributions for all these 28 vectors. The distributions and values ranges on graph and traffic sides turned out to be quite different and all heavy-tailed. Thus, for further experiments, we decided to use robust Spearman(s) rank order correlation measure (and Kendalls(k) Tau, were possible[3]) besides linear correlation(c) coefficient. Users and visits vectors for each fixed graph turned out to be highly mutually correlated: 0.92 (s), 0.79(k), 0.5(c), but because of their different distributions, they represented different information.

Table 1: Graph sizes (nodes) after each preprocessing step (see text for details). $\mathbf{m} = 10^6$, $\mathbf{K} = 10^3$,

step:	g1	g2	g3	g4	g5	g6	g7
0	3m	2.4m	3.1m	2.6m	2.6m	2.6m	2.8m
1(!)	96K	46K	51K	50K	39K	77K	64K
2	78K	37K	40K	39K	31K	63K	51K
(3)4	17K	10K	8.6K	9.5K	7.7K	11K	11K

3. EXPERIMENTAL RESULTS

First, for each graph we measured correlation between in-degree, out-degree on one side vs number of visits and

[1]Obtained by treating each WWW document as a vertex and each hyperlink as a directed edge

[2]Thanks are due to Gemius and PBI, Polish commercial Internet Measuring companies and Netsprint, Polish search engine, for Internet traffic and Web graphs data. Special thanks are due to P.Ejdys, the president of Gemius

[3]Kendalls Tau has high, $O(n^2)$ time complexity

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

Table 2: Correlation between in-degree(i), out-degree(o) on one side vs visits(v) and users(u) on the other side, measured by spearman(s), kendall(k) and linear correlation(c) for each of 7 graphs. All values are in promilles (1/1000)

item	g1	g2	g3	g4	g5	g6	g7
ivs	94	132	100	85	102	107	124
ivk	70	97	73	64	77	80	92
ivc	8	563	30	135	43	136	326
ius	126	151	127	108	135	138	155
iuk	93	111	92	80	100	101	114
iuc	120	495	122	368	148	170	353
ovs	84	41	55	116	77	65	76
ovk	63	30	41	87	58	49	57
ovc	54	12	28	16	22	55	4
ous	45	20	13	80	28	30	40
ouk	34	15	10	60	21	23	30
ouc	57	29	38	12	17	56	17

unique users of Web pages (table 2).

We can see some interesting patterns in the table. Correlation is not strong, but definitely positive[4]. Generally, in-degree is more correlated with traffic data than out-degree. Moreover, in-degree seems to be slightly more correlated with number of users than number of visits. This interesting result can be interpreted as pages having more in-links are more likely to be discovered by Web users. On the other hand, out-degree seems to be slightly more correlated with visits rather than users. Both these positive correlations might be interpreted as users may prefer pages with more out-links as being better hubs. The observed values, however, are not high enough to infer such conclusions with certainty. Note that linear coefficient is much less stable than more robust spearman and kendalls tau measures.

Next series of experiments was devoted to compare PageRank and RBS [2] wrt correlation with the traffic data. PageRank, having one parameter $0 < d < 1$, called *damping* factor[1], is based on a simple model of random surfer [1]. RBS extends this model by including back steps made by surfers[5]. RBS has an additional parameter $0 < b < 1$, $d+b \leq 1$, called *back*, representing probability of back step. When *back* is set to 0, RBS is identical to PageRank.The author thinks that it is very interesting to experimentally check how precisely these two algorithms model the behavior of the **real** surfers of the Web. For each of the 7 graphs, about 60 different[6] RBS score vectors were generated, each for different parameters values[7]. For each such a score vector of each graph its spearman and linear coefficient correlation with corresponding visits and users vector (in the same graph) was computed. Due to space limitations, here we report only the maximal observed correlation values (table 3).

In general, PageRank is observably correlated with traffic data and more than degrees, but the correlation is not strong. Surprisingly, damping values giving the best correlation are often much higher than 0.15 - the value reportedly (e.g. [1]) used in practice. In most cases, maximum correlation was observed for back=0, (i.e.when RBS= PageRank). However, in 8 cases[8] maximum was achieved for non-zero value of back parameter, what means that RBS can better than PageRank model real Web traffic.

The author suspects that RBS will give even better results on denser graphs (e.g. without intra-domain links removal). The methodology presented in this paper may be used for evaluating, comparing or tuning other ranking algorithms. The author also thinks that applying more subtle techniques may reveal more traffic-prediction power in link analysis if even very coarse, global measures described here resulted in quite promising results.

Table 3: Correlation between RBS with $back = 0$ **(i.e. PageRank) vs visits(v) and users(u), measured by spearman(s) and linear correlation(c) for each of 7 graphs. The values represent maximal(m) observed correlation over all checked values of** *damping* **(with** $back = 0$**). Unit is** (1/10000)**. Underlined values indicate cases when higher correlation was obtained for** $back > 0$ **(see the next table). Second half of the first table: corresponding values of damping factor(d) of PageRank, for which the correlation was maximal (the letters refer to next table). Second table: cases, when maximum correlation was higher for** $back > 0$ **(unavailable for PageRank)**

item	g1	g2	g3	g4	g5	g6	g7
vsm	1934	2246	2271	2393	2470	2550	2737
vcm	60	5537	442	733	699	1655	2333
usm	1787	2154	2094	2198	2229	2388	2559
ucm	613	5036	701	1606	1313	1911	2561
vsd	0.7**A**	0.8	0.9	0.7**B**	0.8	0.7	0.8
vcd	0.1**C**	0.1	0.6**D**	0.2	0.2	0.1**E**	0.3
usd	0.6**F**	0.5	0.7	0.4**G**	0.5	0.7	0.6
ucd	0.1	0.1	0.3	0.1	0.07	0.1**H**	0.3

case	PageRank	RBS	back	damping
A	1934	**2088**	**0.25**	0.5
B	2393	**2450**	**0.31**	0.3
C	60	**156**	**0.3**	0.1
D	442	442	**0.1**	0.5
E	1655	**1717**	**0.2**	0.1
F	1787	**1917**	**0.25**	0.3
G	2198	**2251**	**0.35**	0.1
H	1911	**1961**	**0.2**	0.06

4. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th WWW conference*, 1998.

[2] M. Sydow. Random surfer with back step (poster). In *Proceedings of the 13th International WWW Conference, (Alternate Track. Papers and Posters)*, pages 352–353. ACM press, 2004.

[4] All the 3 measures return values ranging from -1 (inverse relation) through 0 (no relationship) to 1 (strongest)

[5] Using "back button" in browsers is very common in surfing

[6] In interesting cases we generated more vectors

[7] More precisely, damping(d) parameter ranged from 0.1 to 0.9 (each 0.1) and back(b) parameter ranged from 0 to 0.3 (each 0.05) what gives 63 possible pairs, but cases with $d + b \geq 1$ must be excluded

[8] I.e. 35% cases for visits and 21% cases for users